WASHC Project
c/o IITA-Nigeria
Oyo Road PMB 5320
Ibadan Nigeria
Tel: (+234) 803 403 5281
Ext. 2744

The WASHC ISFM legacy database

Milestone report 1.6a: Develop ISFM database and ISFM data repository

8/23/2017

E. Jeroen Huising, Samuel A. Mesele and Folasade Adebayo

Supporting Soil Health Consortia in West Africa – Facilitating wider uptake of better adapted ISFM practices with visible positive impacts on rural livelihoods

Report no.: WASHC2017_071

Project Code: 2013 SHP 005]

"Supporting Soil Health Consortia in West Africa – Facilitating wider uptake of better adapted ISFM practices with visible positive impacts on rural livelihoods" is a project funded by the Alliance for a Green Revolution in Africa (AGRA). The project is led and coordinated by the International Institute of Tropical Agriculture (IITA), through the Partnership for Development directorate. Soil Health Consortiums are established in 5 countries, with two consortia for Nigeria, one for the northern region and one for the southern region, and are being hosted and led by our partner institutions:

- CSIR-Soil Research Institute (CSIR-SRI), Ghana
- Institut d'Economie Rurale (IER), Mali
- Institute for Agricultural Research (IAR), Nigeria
- Institute of Agricultural Research and Training (IAR&T), Nigeria
- Institut de l'Environnement et de Recherches Agricoles (INERA), Burkina Faso
- Institut National de la Recherche Agronomique du Niger (INRAN), Niger

Authors of this report

Name: E. Jeroen Huising
Affiliation: IITA-Ibadan
Email: j.huising@cgiar.org

Name: Samuel A. Mesele
Affiliation: FUNAAB
Email: ayodelemesele@hotmail.com

Name: Folasade Adebayo
Affiliation: IITA
Email: f.adebayo@cgiar.org

Database managers with the responsibility for developing the ISFM legacy database for the respective CSHC:

Name: Oyerinde Ganiyu Titilope
Affiliation: Database manager, IAR&T
Email: ganiyuooyerinde@yahoo.com

Name: Andrews Opoku
Affiliation: KNUST, Kumasi
Email: andrewsopoku@yahoo.com

Name: Elijah Ogunsola
Affiliation: IAR, Zaria
Email: ogunsolaelijah@yahoo.com

Name: Maman Garba
Affiliation: INRAN, Niamey
Email: maman_garba@yahoo.fr

Table of Content

Tables

Figures

1. Introduction

This is the milestone report on the activity 1.6a "Develop an ISFM data base and provide access to the data for stakeholders". In the project proposal development of an ISFM database is mentioned in two instances serving two different objectives; one objective is to facilitate the establishment of the CSHC as the repository of ISFM knowledge and the other is to build and strengthen the capacity of CSHC to collate, analyse and synthesize ISFM data and information and develop communication products. Looking at the activities the way they are described in the proposal, they aim to address different aspects of data management. One focusses on data collections and aims to provide harmonized protocols and tools for data collection to evaluate ISFM practices. The other purpose is to improve the capacity of CSHC to generate, synthesize and deliver ISFM information products to its stakeholders. The proposal mentions that the database should accommodate both legacy and newly generated data for the synthesis of ISFM information. It also mentions national and regional databases to be developed and it mentions training needed for the CSHC to use and manage the database. A subsequent and related activity is formulated as to "develop processes and procedures to extract and exchange targeted summary information for the meta-analyses at national level from the national databases towards the development of ISFM recommendations". As such, without being explicit, the proposal intends to cover the full spectrum of aspects of a data management system, from data collection, data storage and handling, data sharing (interoperability) and the analyses of the data and synthesis of results. This is a quite gargantuan task, especially if this must be done from scratch.

In the planning of the activities we had to scale down the ambitions of the project. For this activity, we aimed at providing the building blocks, for the further development of a databases and data management system for accommodating agronomic data sets. We have put efforts in developing the actual ISFM legacy database for the various Country Soil Health Consortia, but implemented in EXCEL, which seemed to be the only achievable option in the end.

The idea behind the development of a database for data from agronomic trials is that it should be able to accommodate data from different sources, from different type of trials and related to different crops. This puts quite different demands on the development of the database and is much more complex than when the database is developed for a particular project, with specific objectives, well defined protocols for the trials and using uniform templates for data collection. Developing a standard database is as difficult as developing standard protocols and as contentious as harmonization and development of uniform tools for the collection, handling and processing of the data. Solutions are found in establishing repositories of data (sets), making sure that these data sets are well documented, or having different databases and making sure that these can 'talk' to each other. The other strategy is to limit the scope of the database, to limit it maybe to one particular crop, or one particular type of trial, for example. The in the end it will come down to the same thing. A successful strategy needs to focus on putting mechanisms is place for sharing of data and assuring the interoperability; that is simply the ability to make use of the data across different platforms and computer systems and for different purposes to some extent.

We have therefore focussed our attention on:

1. Developing a vocabulary of terms used in ISFM and agronomic trials and this includes the specification of concepts and categories in the domain of ISFM that shows the properties and relations between them such that we have a common understanding of how the domain is structured and language to communicate.
2. Developing of meta-data sets, to provide data and information pertinent to the data that allows to properly use and process the data.
3. Development and implementation of the ISFM legacy database in EXCEL for the various CSHC, using the template developed and including quality control
4. Define the data structure, to determine how the data is to be organized, developing the database schema and database itself
5. The data management system and physical implementation of the database, how to organize access to data, etc.

We have made little progress on the latter and we will therefore not report on this.

We have conducted training on data management and analyses and we conducted workshops to train people from the CSHC specifically on the ISFM legacy database as implemented in EXCEL. We have support the CSHC individually in the development of the ISFM legacy database. The training activities have been reported on separately, we do report here on the data that is collected and stored in the various country –based ISFM legacy databases.

Data collection is governed by the specific purpose for which the data is collected. For ISFM legacy data that purpose is defined, namely to collect data on the results from various experiments and trials to allow to synthesize these results to draw conclusion on the effectiveness of ISFM technologies and practices to inform what recommendations on ISFM can be made. A data collection template can be provided for that. For the collection of data from field trials this is different, and dependent on the test crop and cropping system used and on purpose of the trial. It is nevertheless important that the data collection fulfils some minimum standards if it is to be used for integrated analysis and recommendations and guidelines can be provided. We will therefor discuss minimum requirements and will present some example of protocols and of data collections templates to guide the design and data collection for agronomic trials.

Data management issues, issues of sharing data, ‘open access’, interoperability and use of ‘big data’ receive a lot of attention nowadays. This particular work was initiated for the CSHC, but it has relevance in a much broader context. This report is far from complete and certainly not comprehensive in treating the various aspects of the use and management of with ISFM data. However, we hope that this report will stimulate the various institutes that were involved in this project to continue this work, to develop databases and data repositories of ISFM related data and provide open access to the data such that more effective use is made of all the data collected to provide evidence of the effectiveness of ISFM practices and technologies for scaling out and wider uptake of sustainable soil fertility management practices.

2. Vocabulary for agronomic trials and ISFM database

One of the activities in the cadre of the ISFM database development has been the development of a vocabulary on ISFM, or more precise on agronomic trials that are conducted within the context and framework of ISFM. Controlled vocabularies and ontology aim to facilitate data sharing and integration of such data by specifying the concepts and categories in a subject area or domain that shows the properties and relations between them. It helps us to understand how 'the world' is organized (having a common understanding and shared notion) by providing the definitions of terms that are being used that facilitates the communication. The context in which the terms and words are used determines the specific meaning and in this vocabulary gives the meaning of terms in the context of the agronomic trials and experiments related to integrated soil fertility management. Controlled vocabularies and ontologies are used to annotate data, in this case of data emanating from and related to agronomic trials conducted in the field.

ISFM is a concept that is often understood in different ways and therefore needs to be clearly defined. We also need to have a clear idea of what is meant by 'technologies' and 'practices', terms are often used in the context of ISFM, but may be interpreted quite differently. Another example is 'treatment', 'field activity' and 'practice', terms that are easily confused. For the proper structuring of the data it is crucial to have a common understanding of what is meant by 'experiment' and 'trial', and what is understood by 'field', 'block' and 'plot'. Furthermore, terms like 'green manure', 'improved fallow' and 'cover crop' are often use interchanging or as synonyms, and requires clarification.

The vocabulary is presented in Appendix 1. It is very much considered a working document, or document that is in development (living document). The vocabulary was, in first instance, intended to be used in the project to facilitate the sharing of data and the development of the database. It could be adapted and expanded to facilitate the exchange of data and information between projects and other institutions and to widen the scope to include data from other type of trials, related to different crops or cropping systems, or for a different purpose like breeding for example. The research institutes and soil health consortia are encouraged to adopt and expand the vocabulary for their own institutional purposes. There are other ontologies and vocabularies related to the same domain. In the end, we were not able to find one that we could use and that applies to agronomic trials done of purposes of ISFM research. This vocabulary still needs to be harmonized with other vocabularies used in related research fields, like for example the crop ontology developed by the Agricultural Information Management Standards (AIMS) for harmonizing semantics for agronomic data.

3. Metadata sets for agronomic trials on ISFM

The need for metadata sets for the West Africa CSHC

From a scientific perspective, metadata may be characterized as having developed from initially supporting data discovery; to facilitating acquisition, comprehension and utilization of data by humans; and, more recently, to enable automated data discovery, ingestion, processing and analysis via metadata-enabled scientific workflow systems. The continued

conceptual and operational developments in metadata required to support comprehensive automated scientific workflow systems portend many challenges and opportunities.
Numerous discussions on data management associated with the study of climate change, studies on biodiversity and sustainable development have highlighted the need for accepted protocols to assist scientists with preserving important data sets and providing guidelines for the supporting documentation that is necessary to interpret the data (National Research Council 1995). Similarly, for the advancement of agronomic research, to develop more sustainable soil and crop management practices, better use of existing agronomic through metadata-enables exchanged of data and information. Although geospatial and ecological metadata standards have been developed and widely endorsed by the geographical and ecological science communities respectively, such standards do not yet exist for agronomic field experiments. Agronomic data and information is poorly managed, hardly available and what is available difficult to discover. It requires the various institutions and organisations to adopt metadata standards to allow for the sharing and use of agronomic trial data. The West Africa Soil Health Consortia (WASHC) project aiming to make use of existing data from agronomic trials conducted within their own country specifically, embarked on the development of metadata sets to document the data retrieved from different sources. To be able to make recommendations on ISFM based on results from various trials, you need to be able to interpret results from these trials correctly and for that you need to know the provenance of the data, the way the data was collected, what the data refers to, the treatment it corresponds to etc.

Definition of metadata

Metadata may be defined as information about data — i.e., the information required to understand data, including data set contents, context, quality, structure, and accessibility (Michener, 2006). It is descriptive information about a particular data set, object, or resource, including how it is formatted. Concisely, metadata describe the “who, what, when, where, and how” about every aspect of the data.
A metadata standard on the other hand, would normally support a number of defined functions, and will specify elements which make these possible. They are requirements which are intended to establish a common understanding of the meaning of the data, to ensure correct and proper use and interpretation of the data by its owners and users.
The general objectives for metadata implementation include facilitating:

- Identification, discovery and acquisition of data for a specific theme, period in time, and geographical location
- Determination of the suitability of data for meeting a specific objective; and
- Data processing, analysis, and modeling.

We make no distinction between the metadata set for ISFM legacy data or for ISFM data sets. With the term ‘legacy data’ we generally refer to data that has been published or otherwise reported. This generally is processed or analysed data, however increasingly we find people publishing the data itself. In providing data on the experiment or trial, or the provenance of the data, there is no principle difference between whether this concerns published results from the trial or experiment (processed data), or whether it concerns the data from the trial itself. The metadata sets we propose applies for both the published articles as well as data sets, though there are some complications when it concerned processed data as discussed below.

Generally, 3 categories of metadata can be distinguished, namely: administrative, descriptive, and structural metadata:

- Class 1 or Administrative metadata: This relates to the source of the data and may include descriptors like citation of the article, project reference, data owner or contact person, time when created, update history if it concerns data sets, access restrictions, reference to associated documents, etc.

- Class 2 or Descriptive metadata which provides information on the research context and which may provide descriptors for the experiment, like the hypothesis tested, or experimental feature, the experimental factors considered, the type of trial, time when the experiment was conducted, geographical location of the experiment, test crops considered, etc. The purpose of these descriptors is to alert potential secondary users to the existence of the data sets and thus to improve the discoverability of data sets that falls within specified temporal spatial and thematic domains.
  Besides the descriptive metadata to provide information on the research context, we also have descriptive metadata that provides information on the treatments, which is considered to be part of the trial or experimental design. This purpose of this type of metadata is to assist users in bringing similar resources together for integrated analyses and meta-analyses for a specific research question.
  A third sub-category of the descriptive metadata is what we refer to as the 'technical metadata'. This provides information on type of trial in terms of the statistical design, the sampling design referring to how the locations for the trials were selected, in it specifies the data collection method and method of measurement. It may specify how is dealt with missing values and aspects of data quality and integrity. The purpose of this type of meta-data is to assist the user in deciding whether the data can be used and for what type of meta-analyses.

- Class 3 or Structural metadata: metadata about the containers of the data and how the data is structured and organized. It describes the data types and relationships between the digital materials. This includes the 'data level metadata' which refers to the description of the attributes, provides information of the units in which the data is expressed and information on the domain values and other. This is the data dictionary. This type of metadata provides the user with information on how the data can be used, whether conversions are needed and other.

The metadata standard for ISFM data for the West Africa CSHC and its use.

The vocabulary, discussed in the previous section and data dictionaries are tools that help to standardize the metadata structure. The complete list of metadata descriptors (the attributes) with their description or definition of the metadata set used for the ISFM legacy data is presented in Appendix 2. This is the metadata standard adopted by the various CSHC. It is not a very elaborate list. The information required is like the information that is required to be provided when submitting a data set for publication. It is information that in principle you should be able to extract from a well written and comprehensive publication on an agronomic experiment or trial. We consider this metadata standard to represent the minimum data requirements for the metadata to be able to fulfil its function, namely to 1)

facilitate the identification and discovery of agronomic data relevant to ISFM, 2) to provide information where the data can be sourced, 3) to facilitate the selection of data for answering specific research question and suited for particular type of analysis, and 4) information assisting in the correct use of the data.

For the WASHC project the principle use of the metadata set is to describe the data that is obtained from published articles. To correctly use this metadata standard, it is key to understanding what the data in the research article or report actually refers to. In a research paper the descriptive statistical data is generally presented, and the data may have been aggregated in different ways. Results/data may be presented for individual trials, or for several trials combined trials (e.g. if trials are replicated at different locations or replicated over different seasons); the data may refer to the results obtained for individual treatment factors (even though the treatments itself are a combination of treatment factors) or for the individual treatments (i.e. the specific combination of treatment factors). It depends on how the data is presented. The metadata needs to provide information on the data in the way it is presented (and captured in the database) rather than how the trial is designed. This has implications for how the treatments are described, but will also have consequences for what is considered as metadata and how the data is structured. For example, when data for individual trials is captured, the coordinates of field location can be specified (and can be part of the metadata), but if the data refers to a set of trials at different locations this does not apply and the location needs to be defined in another way. There are several solutions for this kind of problems. Important is that the metadata structured such that these different instances can be adequately accommodated. For the use of the metadata set it is important to adhere to the strict definition we have adopted for what an experiment, experimental feature, trial and trial location, treatment, etc. is.

The metadata aims to describe the experiment, that may relate to one or several trials of a particular design, conducted at one or several locations during one or more seasons. The experiment is therefore the entity for which data is provided. The source of the data is generally a publication or report and we assume the publication refers to one experiment. However, the experiment may have different research aspects (related to the different research hypotheses) that that may reported on in the same or in different articles. As indicated in the previous paragraph it depends on how the data is presented in that particular article whether we need to distinguish between these different research aspects in the descriptive metadata. To accommodate that we have introduced the concept of experimental component, such that data related to the various research aspect is presented separately while referring to the same data source. If there are several articles reporting on different aspect of the same research/experiment we will have the different sources of the data listed in the administrative metadata. It does not make a fundamental difference, whether the different research aspects are treated as different components of the same experiment or as different experiments. It has to do with choices on how to structure the data rather.

The above is illustrated by the way how the data was captured from an experiment reported on by Azeez etal (2009). In this experiment, the effect of weeding intensity in combination with different N application rates and different maize varieties on maize yield was tested in multi-factorial trials conducted in two locations. In the article the data for the two locations were presented separately, the data (results) for the weeding effect (three intensities) is

presented separate from the results of the N application (4 rates of N-application) and separate from the results for the different maize varieties (4 genotypes tested). Each treatment was replicated three time. The results for the weeding effect is aggregated for the N-application rates and maize varieties, as the results for the N-application rates and the results for the maize varieties was also aggregated. For the data on the weeding intensity the corresponding number of observations in 28 (that is, 3 for the number of replicates times 4 for the N-application rates, times 4 for the maize varieties). In the databases this was accommodated by identifying three experimental components, one for each experimental factor investigated. Had the original data been presented, with the descriptive statistics for each treatment combination this could have been captured as one experiment with one component only. In that case, the number of observations would be 3, related to the number of replications.

We have made extensive use of value list to control the entry of values for the various attributes. This is intended for quality control and preventing errors in the data entry and at the same time allowing for specific data values only facilitates the search for data in the database. We have applied this wherever this is opportune and does not provide restrictions for any type of trial or experiment. These value list especially apply to the various attributes of the treatment, which are easy to categorize (like tillage operations for example). The trade-off is the level of detail that you can allow. Further testing of the metadata set will tell whether the current standard serves it purpose or whether adjustments have to be made. The value list for the various attributes in the metadata set are presented in Appendix 3.

The ISFM metadata set compared to other meta-data standards

The ISFM metadata is compatible with many other metadata standards and metadata sets. We need to consider, though, that the ISFM data sets (and therefore the metadata) relate to a specific activity, namely agronomic trials. In the metadata standard for the AfricaRISING project, the data set can relate to a wide range of activities, from household surveys to lab experiments, for example. Similarly, the CG core metadata allows for data sets related to a wide range of subject matters or research areas. This allows the ISFM metadata set to be more specific on the description of the experiment or study, specifically the description of the treatment. We consider the treatment description to be part of the metadata, which will allow the search for specific data, which will facilitate the reuse of the data. None of the other metadata standards provides the possibility to describe the trial design, including the treatments, not even not the metadata set developed by IPNI intended for ISFM legacy data collection.

Otherwise, the ISFM metadata set contains elements that in the other metadata standards are considered to be part of the minimum data set; that is, they belong to fields that are required to be filled. The other metadata sets generally provide opportunities for a more comprehensive documentation on the project and activities. For example, the CG core metadata allows to list all the contributors to the data set; in the AfricaRISING metadata sets all partners in the project can be listed. In the ISFM metadata set there are no provisions made to enter such information, because that information is not considered essential and can generally easily be obtained through other means if needed. Very elaborate metadata entry

templates will not entice people to fill the forms, and for that we want to keep it as simple as possible.

Metadata template for the ISFM trial data

For entering the metadata a template has been developed, implemented in Excel. It is structured such that we have separate sheets for the various types of metadata: a sheet for the administrative metadata that provides information on the source of the data, one sheet for the descriptive metadata concerning the general design of the experiment and trials, and one sheet for the description of the treatments. This is different from most metadata entry templates that have one data sheet for entering all the metadata, in which the data elements are organized in one sheet per row rather than per column. The data structured in this way reflect the database design in which the individual sheets resembles the tables in the database. In this way, it is also easy to enter the data. In case you have a data file repository, in which you have spate metadata files attached to each data file, the data should be structured differently.

The country soil health consortia can expand the metadata set according to their specific needs and scope of the data that they want to include, as long as it contains the elements in the current metadata set, which reflect the minimum required data. It should always be possible to map the elements of their metadata base to the elements of the current ISFM meta data set. We have provided the various metadata sets developed for other projects, like the OFRA project for their reference, if they want to develop the metadata sets for specific experiments and activities.

4. ISFM legacy database of the four CSHC

Design of the ISFM legacy database

The ISFM legacy database contains legacy data from ISFM trials that is generally obtained from published articles, reports and other. The ISFM legacy database aims to make is possible to compare and synthesize results from various trials. The data that is gathered and entered in the database therefore relates to common recorded variables like yield, based on which such comparative analysis can be done, rather than data on variables that are quite specific and relevant to that specific trial or experiment only. For example, the number of plants harvested, or data on plant growth, leaf area index or other is relevant to the specific experiment only and therefore not included on legacy database. The data in the ISFM legacy database refers to processed data (that is descriptive statistics) on crop response to treatments, rather than the original recorded trial data that relates to data on crop performance for each individual plot. The ISFM legacy database is not a database in the sense of a data file repository, in which the files contain data from one or more trials conducted under one specific experiment. However, the data structure of the ISFM legacy database can be used, with some adaptation, to capture data recorded from individual plots/treatments of the trial. The statistical information captured in this database is the average yield and the standard deviation, standard error of the means (SEM), or the Least Significant Difference (LSD) associated with that yield assessment. The number of observations is an important parameter to be listed to allow for further quantitative analysis and comparison of results from various trials or experiments. This structure still allows entering data on crop response

to treatment for individual plots, as an exceptional case in which the number of observations is one (1) and where the other statistical variables (standard deviation, or standard error of the mean) for that record will be 'null'. The database accommodates both data from individual trials (conducted at one particular location and in one particular season) with or without replication of the treatments, and from a set of trials (either trials conducted at multiple locations or in several seasons). The most detailed data available should to be entered in the database. That is, if data for the individual trials is available that data should be entered.

For each type of crop there is one particular table for entering the legacy data, because yield is referring to the different plant parts and measured in different ways for the various crops and is therefore also interpreted in a different way. So far in the database distinction is made between banana as a crop (musa species), grain crops (including grain legumes), root and tuber crops and vegetables. For banana (whether dessert or cooking banana or plantain), the yield is difficult to establish and depends on the time between subsequent bunches are produced by the same mat, the plant density and the bunch weight. Comparison of the yield between experiments can only be meaningfully done if data on all these parameters is provided. The effect of various soil fertility treatments is expected on bunch weight mainly. Therefore, provision is made in the database for entering data on bunch weight, number of hands per bunch and number of fingers per bunch, only. For cereals and grain legumes provision is made to enter data for the grain yield and stover yield, and for the 100- or 1000-grain weight, depending on the type of grain. Provision is made to enter data for fresh weight and dry weight of the grain and stover. The provision for entering dry weight is not made for root and tuber crops, because the yield is generally provided in fresh weight. This means that for those crops and experiments where the yield is provided in dry weight or as flour the conversion needs to be made to fresh weight using commonly accepted or measured conversion rates from fresh weight to dry weight or by using established milling extraction rates. The same applies to the vegetables, where provision is made only for entering fresh weight. For grain legumes, there is an option to include nodule number and nodule weight, since this is often used as an indicator for nitrogen fixation in legume crops, which is the subject of many studies. But this may be too specific and if to cater for this option it is best done by define a separate table for this purpose. The same applies to data from pot experiments, in which 'seed weight' and 'shoot weight' have a separate meaning. Treatments in pot experiments are also described differently from the way treatments are described in field experiments. Therefore, separate and specific provisions should be made for storing of data from pot experiments. For root and tuber crops (e.g. cassava, yam, potato) the data for tuber yield, the commercial tuber yield and the non-commercial tuber yield can be recorded, by entering the weight percentage of the commercial sized tubers and the average number of commercial sized tubers per plant (to contrast with the average number of tubers per plant). The distinction between commercial and non-commercial tubers is made based on the size of the tubers or roots. If data is presented for both categories, the criterion used to distinguish between commercial and non-commercial sized tuber or roots should be specified and recorded as part of the metadata. Finally, for vegetables, the shoot yield, fruit yield or seed yield is recorded, depending on what the harvested component is. Vegetables like carrots are classified as root crop and data needs to be entered in the corresponding table for root and tuber crops.

The general characteristics of the field where the trial is conducted are described in the 'field' table. These properties refer to the location, site characteristics, soil properties, land use and land management history, in agreement with the information generally provided in the research article. Distinction is made between properties of a more permanent character and those of a more temporary character. The more permanent properties for site characteristics are topographic positon and slope. Land use history is captured as the 'years cultivated' (years since deforestation or that the land is reclaimed for agricultural use) and as 'fallow type' and 'fallow duration'. No provision has been made to record past management practices, like tillage operation, fertilizer application, etc. Many consider this important information to be able to evaluate the outcome of an experiment. However, this information is seldom provided and therefore no provision is made to store this kind of information in the current ISFM legacy database.

For the soil the more permanent properties are the soil classification, textural class and soil depth. Other soil properties, properties like those determined by sample analysis (chemical, physical, and biological properties, like, nutrient content, soil organic carbon, CEC and ECEC, Base saturation, bulk density and other) are stored in a separate table. These are also the properties that an experiment often tries to influence or change (improving soil condition or quality). Entering this data in a separate table allows multiple records to be entered for the same field/location. This caters for the situation in which the soil properties are determined before and after the trial, for example. Also soil properties are often determined at the start of each experiment, even if this is conducted on the same field. To interpret the results for the soil analyses the method of extraction and method of measurement needs to be known. This information (metadata) is entered as free text in one data field in technical metadata table for all elements that were analysed together, rather than for each element of method of analyses separately.

Finally, there is a table for water supply data. For rainfed systems, this refers to the cumulative rainfall from time of planting to the time of harvest. This is the only measure that applies to the various type of crops. Many articles provide information on long term averages on cumulative seasonal rainfall or rainfall distribution statistics, rather than rainfall being actually measured in the field. This data may be entered as long as the source of the data and method of collection (what the data represents) is detailed in the metadata descriptor for this purpose. In case of irrigation (or if other methods of additional water supply are applied) the amount of water supplied through irrigation should be added.

The entities and their relationship in the ISMF legacy DB

The following entities have been defined that apply to any experiment or study involving agronomic field trials. How the entities are defined determines to a large extent how the data in the database is structured. The ISFM legacy database needs to accommodate data from various research papers in which results from the trials may be aggregated to different levels or degrees, and this has implications for how the data is structured and therefor how the entities and relationships between these entities are defined and how these are implemented in the database. This may be different from the databases designed for a specific experiment. For example, the crop yield data may refer to the treatment in a specific plot, the treatment in a specific trial or to the treatment in the experiment implemented in several trials.

Experiment: With experiment we refer to one or more trials conducted at one or more locations within a specified region, at one or several times (seasons) within a specified period, aiming to give answer to one or more research questions within a specific problem domain and following one particular design for the trials. The design specifies the layout of the plots, the treatments and the statistical design of the trial, as well as how the locations for the trials are selected in case trials are conducted at multiple locations. Consequently, if an experiment with the same purpose and design is conducted in a region differently from the specified region or conducted in a period that is different from the specified period it is considered a separate experiment. There can be several experiments in the same study, for example if experiments are conducted in several countries.

Trial: The 'trial' is defined as one instance of an agronomy experiment; A trial is implemented on a particular field and at a particular time (cropping season or year), according to the design as it is defined by the experiment the trial belongs to (that includes the specification of the various treatments). For a complete block design and for which the data is presented for each block separately, the individual blocks can be designated as separate trials (each with their individual identifier) referring to the same experiment having the same design (referring to the set of treatments), being conducted on the same field and established in the same year. A trial that is replicated in the next season (or year) is recognised as a separate trial (as identified by the year of establishment) but with the same experiment reference (the same design), and is referenced to the same field (same field-ID). In study where crop rotational effects are investigated in subsequent seasons, for example in a split-plot design, we still have separate trials that are distinguished by the season or year of implementation that are implemented in the same field (same field-ID), but regarding different treatments. This requires that the relationship between the subsequent treatments in the same plot needs to be established, for which provisions need to be made that are not in the current database design.

Field: The field refers to a particular piece of land where a trial is conducted; the field is defined by its boundaries. The same field may be used for trials in subsequent seasons as part of the same experiment or as part of another experiment. Each field has a specific location and specific site characteristics and soil characteristics. Also, each field has its own history in terms of land use and land management.

Treatment: Treatment is defined by the specific combination of agronomic practices applied at a particular plot in the experimental field. The agronomic practices refer to land preparation, crop management, the use of agricultural inputs, use of organic inputs, weed management, pest and diseases management and crop water management. Typically, in an experiment one or more of these practices vary as per design, which then represent the experimental factors. Like the 'trial' refers to the instance of a field experiment, the location and time of the trial, so does the 'treatment' refers to the instance of a combination of management practices (as per treatment design) at a particular location (plot) and time (season) as part of a particular trial. Distinction is made between the 'treatment design', referring to how the practices in that particular 'treatment' are defined and the implementation of the treatment; that is the occurrence of the 'treatment' at a particular locations and time.

Plot: The plot is that part of the experimental field in which one particular treatment is implemented. Each plot is assigned to a specific treatment, but you can have the same treatment on different plots (if the treatments are replicated). All observations related to crop growth and yield are done at plot level and the plot forms the smallest unit of observations on the trial field (though measurements may be done on individual plants within the plot, but those observations will not be recorded in the ISFM legacy database). The specific plot does not need to be identified, unless there is specific information associated with the specific plot (e.g. in case soil properties are recorded for the specific plot rather than for the field or when yield data is provided for different plots where the same treatment is implemented). Otherwise the treatment identifier, together with the identifier for the trial will suffice as reference for the yield data record. If the yield data refers to treatment effect on specific plots, the treatment identifier with a symbol (e.g. a letter suffix) to denote the plot and to distinguish between one and the other plot with the same treatment will suffice.

Crop: With 'crop' we refer to the crop in the 'field' (that is the instance of the crop grown in a particular field or plot), rather than as crop as an abstract term. This may refer to a single crop or a crop grown in crop combination with other crops, at the same time or in relay, and for which data on crop performance (crop growth and yield) is presented in the database. In the case of the ISFM legacy DB the crop yield data refers to the treatment, whether implemented at a particular plot, in a trial or particular field, or in a particular experiment (depending on whether the treatment is replicated in one trial or replicated in several trials of the experiment). Alternative to 'crop' we could call it 'crop response' or 'crop performance'.

Soil: Soil refers to the particular type of soil as determined by the soil texture, chemical, physical and biological characteristics and other characteristics as indicated by its soil classification. The 'soil' may refer to the soil of the plot or of the experimental field or other spatial unit.

Implementation arrangements

In the current design for the ISFM legacy database the identifier for the 'Trial' is made up of the identifier of the experiment, together with the identifier of the field and the specification of the season or year the trial is established. The results of the experiment are often given for each trial separately, but may also be given for several trials combined (for the experiment conduced at different locations, or in several seasons), in which case the field is not identified (enter "99" for field ID), nor can the year be specified (enter "9999" for the year of trial establishment), in the table where the yield data is presented. Results are presented for the different 'treatments' whether for individual trials or for a several trials combined. Only in the exceptional case that data is provided for individual plots, a symbol needs to be added to the treatment identifier to denote the replication and herewith the individual plot is also identified. In the original trial data, results are provided for each plot individually, but this this does not apply to the ISFM legacy database that in general presents statistical data and a reference to the plot number is therefore not needed. For a database to contain the original trial data adjustment to the design should be made, which could be done by adding a separate column for 'treatment replication'.

The database is implemented in EXCEL, which is not a database management systems and therefore has its limitations in managing the data. The possibilities for validating the data upon entry are limited and so are the possibilities for querying the data base. We have tried to maintain the same structure in EXCEL as we would have adopted for a relation database, meaning that we have kept in separate sheets what we would otherwise kept in separate tables. EXCEL however does not allow us to manage and maintain the relationships between the various entities. Retrieval of data making use of these relationships (linkages between the records in the various tables) must be done by hand. An EXCEL database is also not very efficient for providing the access to the data online. Because of the above we aim to implement the database in a relational database and we have developed a prototype implemented in MariaDB, which however needs to be further tested.

Furthermore, we have currently four (4) EXCEL workbooks, each containing data gathered by the respective CSHCs from sources that are relevant to their respective countries (that is, data from trials and experiments conducted in their countries or specific region). The number of data sources captured in the respective databases is still limited and it is for the CSHCs to decide whether it is worthwhile to invest in developing a relation database for the ISFM legacy data.

The way it is currently implemented in EXCEL makes it relatively straightforward to convert this into a relation database. Though normalisation of the database would require tables to be split and you would probably include several tables just to model and establish the links/relationships between the entities. Also, the value list wold appear as separate tables. The tables with the definition of all the data fields in the table are presented in Appendix 4. The data fields described include those fields that serve as 'foreign key' that establish the link to the other tables. If we include 4 tables that contain the metadata, that is also part of the database, then the total number of tables in the database is 11.

Table 1 Tables contained in the ISFM legacy database

Table Name	**No Attributes (incl. foreign keys)**
Field	15 fields defined
Soil characteristics	26 fields defined
Grain yield	27 fields defined
Root&Tuber yield	21 fields defined
Banana yield	20 field defined
Vegetable yield	22 fields defined
Water supply	5 fields defined
Metadata tables	**No attributes (excluding foreign keys)**
Treatment	32 fields defined
Experiment design	3 fields defined
Experiment description	18 fields defined
Data source experiment	11 field defined

Quality control and assurance

The ISFM legacy database contains data extracted from reports and publications that have been reviewed before publication (articles in peer reviewed journals, MSc and PhD thesis, official reports). We (have to) assume that this gives some kind of assurance of the quality of the data presented, which is expected to be free of gross errors. However, we know little about the quality of the original data that is used in the calculations. If the article contains information on the quality control procedures applied, this should be captured in the metadata (and provision is made for that in the metadata set).
Otherwise the quality of the data depends on how accurate the data is extracted and copied from the article. For tabular data, the data is mostly extracted 'by hand'. Which gives opportunity for error. For extraction of data contained in graphs or charts a digitizing tool can be used to capture the values, but the use of that tool that may cause error. Errors may arise in linking the yield data to the specific treatment, and the correct description of the treatment. The design of the trial is generally described in the methods section and this does not always relate directly to the way the data is presented. It is, then, also not always clear what the associated number of observations is, for example. It requires experience, skills and expertise to read the articles correctly and extract the right information, and proper training is therefore required. Rules and procedures for checking the data and extracting the relevant information are not easy to give, and for the same reason it is difficult to come up with standard quality control procedures. It would, therefore, be good practice to have the data that is entered checked by a second person, which would require the have the report or research paper available. It would therefore be good to keep the electronic copies of the report or papers in a repository. It is more and more becoming a habit to also make the supporting data for research paper available, and these documents should be referenced as well in the database and made available for possible future reference. There should be someone with the final responsibility to declare the data accurate and reliable.

To preserve the quality of the data and integrity of the database extensive use is made of look-up tables (or value lists) to enter data automatically by copying the data from related tables where applicable. This reduces the risk of error and improves the consistency of the database. Where applicable data is entered by calculation based on data from data fields in the same or other tables. These data fields are not supposed to be entered manually or to be imported from external sources, though it is possible to overwrite the calculated value. So far, we have not developed rules for the validation of the data when it is entered. This could apply for example to data that is to be within specified value ranges (e.g. for yield data) to detect typos. As mentioned this is not implemented that easy in EXCEL but should be recommended when converting to a relational database. Simple rules that empty cells are not allowed should be applied and adhered to. There should also be rules on the use of majuscules to improve consistency in the way the data is entered and to avoid errors in searching the database.

Country Soil Health Consortium ISFM legacy database statistics

There are 4 CSHC that have worked on the legacy database and that have used the data collection template developed in EXCEL to enter the legacy data, i.e. data extracted from the reports or publications.

The Nigeria CSHC Southern node is divided in two regions: The South-South and _South-West. It comprises of the following states: Oyo, Osun, Ogun, Ekiti, Ondo, Lagos, Edo, Delta, Akwa Ibom, and Enugu, and the data they are gathering are from trials conducted in those regions. The Nigeria South soil health consortium legacy database consist of 55 datasets from experiments/trials conducted in the period from 1995-2014 (20 years). There are 2 experiments that concern intercropping. The database contains a total of 441 treatment combinations, of which 111 include the used different types of manure to improve soil fertility. Poultry based manure and soil additives were the most used with 78 treatments involving the use of this organic resource. The use of human faeces and urine is among the manure types that is investigated. The trials for which the data is captured cover 23 crops, with maize being the most dominant test crop (Figure 1). Thirty-nine (39) treatments evaluated the use of compost and 29 evaluated the effects of different organic fertilizers, while combined the used of organic and inorganic fertilizer was evaluated in 15 treatments. The use of improved crop varieties was evaluated in 240 treatments.

We do not know whether the list of articles from which the data is extracted is exhaustive, but the spread in crops tested is noteworthy and we observe that for most crops there are data available form only a limited number of trials. Given that the experimental feature from one experiment to the other will differ for one particular crop, it will be unlikely that the data can be used for comparative analyses, even if several experiments have been done using the same crop, like for maize. The conclusion is that far more systematic research has to be done on specific ISFM practices for specific crops to get any kind of conclusive results.

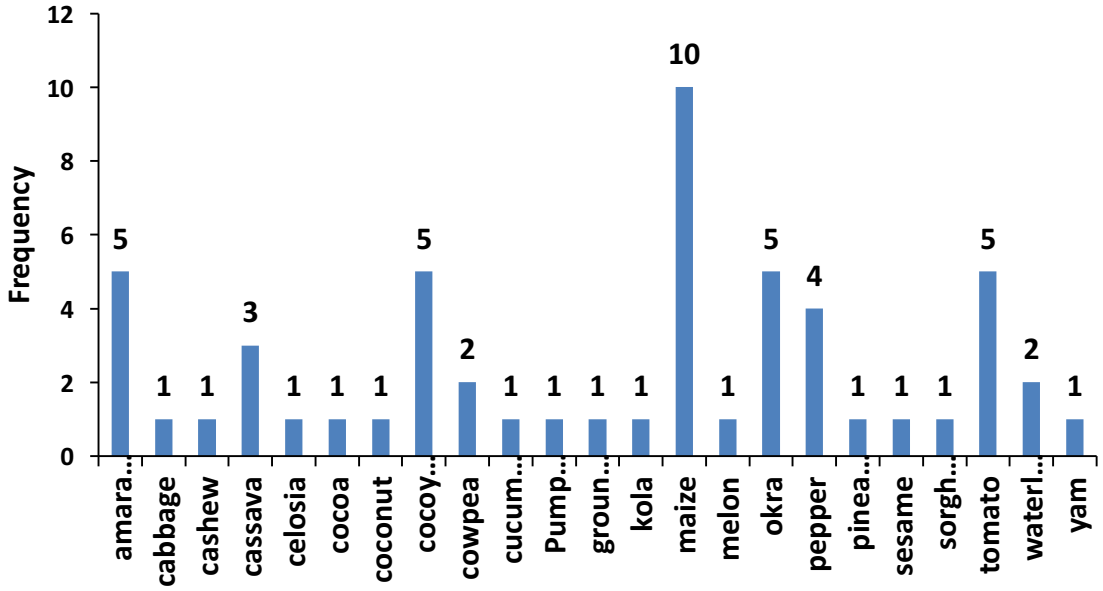

Figure 1. Distribution of the number of data set per crop contained in the database

The Nigeria Northern node oversees the collection of agronomy trial data from the northern region. This comprises of Kaduna, Kano, Katsina, Sokoto, and Niger State amongst others.
The Nigeria northern node ISFM legacy database consist of 61 datasets from 1990 to 2014 with only 1 trial on intercropping. A total number of 444 treatments combinations were identified of which 120 used different types of manure to improve soil fertility. Poultry base manure and soil amendments were the most used in the trials with 54 treatment entries. A

total number of 19 crops were captured in the database with maize been the major crop (Figure 2). In 167 of the treatments improved varieties were used.

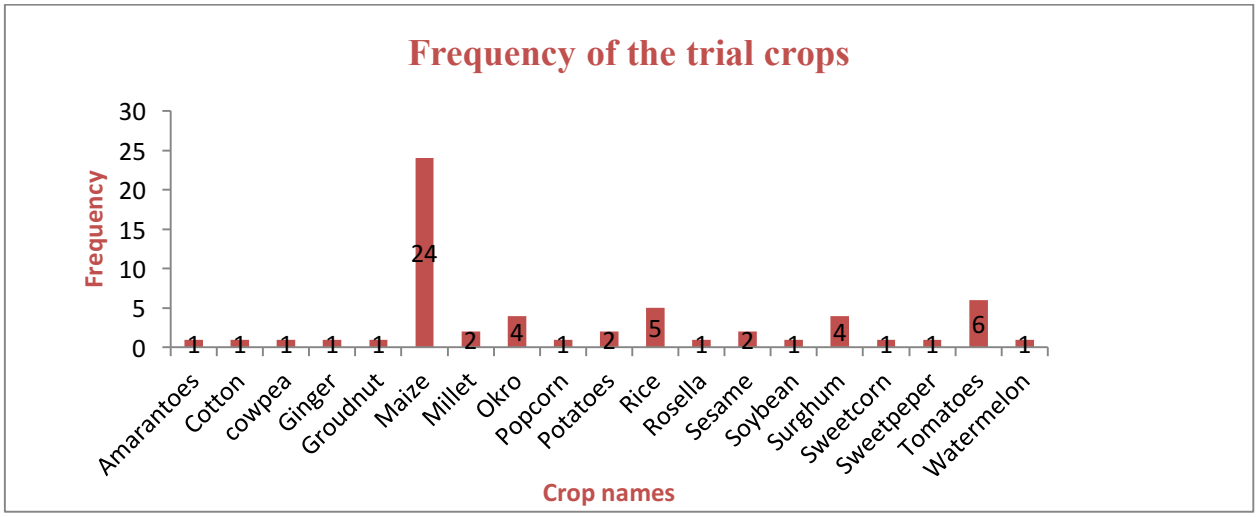

Figure 2 Frequency of data sets for each crop captured in the ISFM legacy database of Nigeria SHC northern node

As similar conclusion can be drawn regarding the number of trials for which data is gathered, namely that these are not enough to allow for any conclusive results on the effect of ISFM practices on any of the crops. There must have been many more trials conducted, but these have probably not been reported on, if at all the data has been analysed.

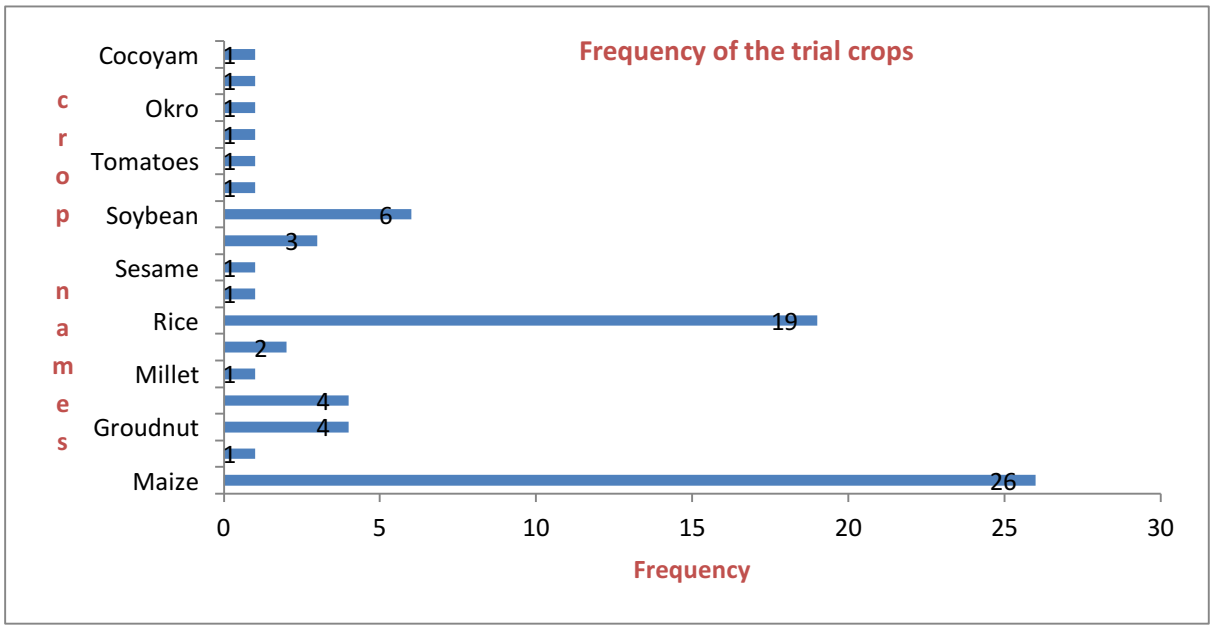

Figure 3 Bar chart of the frequency distribution of data for the various crops in the Ghana ISFM legacy database

The Ghana Soil Health Consortium database consists of 73 datasets collected from trials conducted in the period from 1992 to 2014 (22 years) with four studies reporting on intercropping system involving either of the following crops: cassava, legume, sorghum and

soybean crops. A total number of 875 treatments combinations were identified of which 108 used different types of manure to improve soil fertility. Treatments with cow dung or poultry manure, with or without the use of crop residue, were the most frequent with 89 treatment entries. A total number of 17 crops was captured in the database with maize being the dominant trials (Figure 3). The use of local varieties was evaluated in 205 treatments while the remaining treatments was based on improved varieties.

For Ghana we see the number of data sets for maize and rice considerably higher than in the other databases, and it might be worthwhile to investigate whether there are some general patterns emerging from these trials related for example to the use of organic resources and to the response to use of inorganic fertilizer.

The Niger Soil Health Consortium legacy database consist of 66 datasets covering a period from 1986 to 2014 with 12 data sets from trials on intercropping or crop rotation involving groundnut or cowpea. A total number of 367 treatment combinations were identified. Hundred (100) trials used different types of manure for soil fertility management. A total number of 4 crops was identified in the database with Millet being the most dominant. There is a lot on missing information in this database, which results in a number of trials for which the test crop is not specified. An effort should be made to collect the missing information such that further synthetic analysis can be done. Having databases developed by the Mali and Burkina Faso CSHCs would have been useful to increase the number of data sets available for sorghum and other crops currently under reported. Figure 4 below shows the bar chart of the result from the analysis.

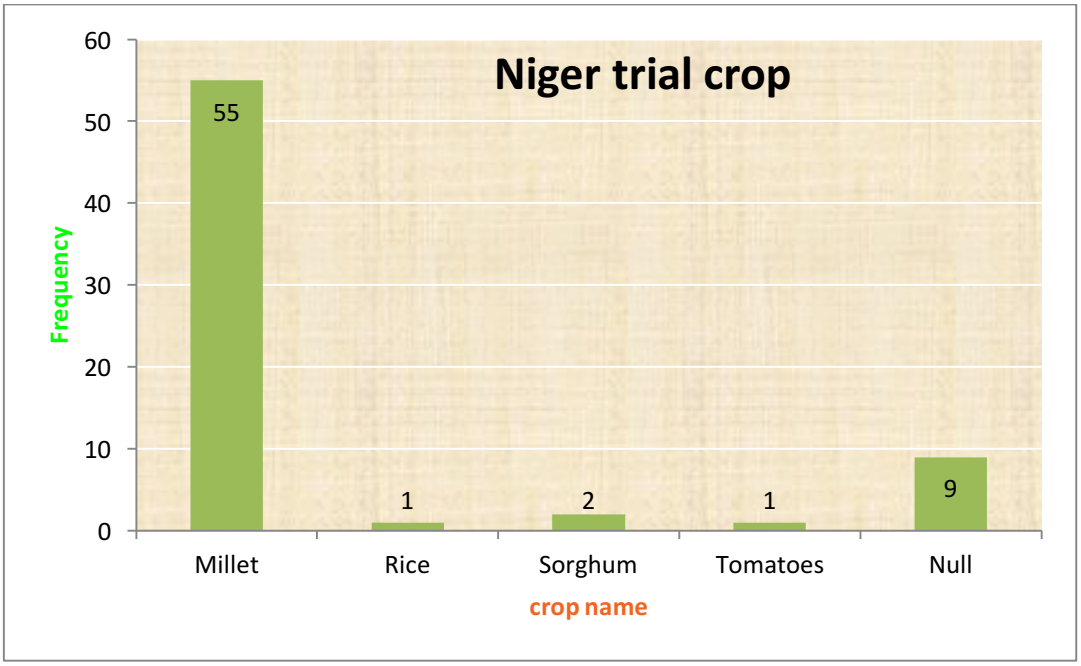

Figure 4 Number of data sets available for the various crops in the ISFM legacy database of Niger

Currently the four databases are not integrated into one regional database. This is because the four databases are slightly different, though they have the same basic structure. It would certainly be useful to consider the data sets from the various databases together for further

analysis. For maize, bringing the data together would result in 60 data sets, for example, which would probably be enough to do some meaningful meta-analyses, if the data would allow. At present this information could be easily extracted from the four databases, without the need for the databases to be integrated.

5. Data collection and minimum data sets

The question of the essential data set (sometimes called minimum data set) is difficult to answer and what is considered as essential depends on the context for which the data is being used. Data that is essential for the ISFM legacy database is be different than the data considered essential for a trial or a specific experiment.

Data that is essential for the ISFM legacy database needs to be entered. The data fields should therefore be marked as 'mandatory' or 'not null'. The criteria for deciding whether data is essential or not should be based on the consideration whether the other data in the data set (*viz*. the yield data in this case) would still be relevant and fit for use if that data element is not provided. This depends on the type of analyses to which the data will be subjected which is generally not known. Therefore, one should be very restrictive in declaring attribute data essential, in order not to discard any data sets that could still be used for a particular application that was maybe not directly foreseen. The idea would rather be to use the availability of certain attribute data as part of the search criteria for finding data for a specific application. For example, soil data might not be essential for comparative analyses of results from experiments that look at the effect of tillage operations or weeding practices. A similar question can be raised with respect to the need for data on a control treatment. For many experiments, the control treatment provides the basis, based on which the results of other treatments are evaluated, and which subsequently makes it possible to compare between results from various experiments (based on relative response characteristics rather than absolute values. However, for certain type of experiments it is difficult to define what a control treatment should be.

In any database, the primary key and secondary, or foreign, keys are essential because they are needed to maintain the referential integrity of the database. The primary and foreign keys are identified in the data dictionary in Appendix 3 and will not be further discussed. For the ISFM legacy database we consider the data listed in table 2 below to be essential. The data items often belong to the metadata. This is because these are attribute data that are used to search the database or otherwise provides information that is needed to assure the correct use of the data.

Table 2 Mandatory data fields in the ISFM legacy database (metadata)

Data field/descriptor	Explanation
Source/citation	To make sure there is a trustworthy / reputable source for the data.
Experimental factors	One of the main criteria used for searching the database probably
Experimental feature	What is investigated will provide an important criterion for searching the database
Crop and cropping system (second crop name)	Probably the main criterion used for searching the database; the yield table also contains the crop name, which is likewise a

	mandatory data field, but that will not tell whether the crop is part of an intercropping system on relay cropping system.
Year	Used in the metadata set as well as in the yield table; Often important search criteria, but also important reference information for assessing auxiliary data like rainfall statistics
Location	Information often used for searching the database and provides important reference to auxiliary data like soil data, climate data or other
Data collections and method of analyses	Important reference information required for proper interpretation of the data. E.g., the results from soil analyses cannot be properly interpreted if the method of analysis is not known
All data fields in the Treatment table: Crop variety, Plant density, Nutrient application rate, etc.	All the data fields in the 'treatment table' should not be left empty ('not null') because without the description of the treatment the crop response data becomes meaningless. For all practices data can in principle be entered, even if the practice is not applied. For example, if no manure is applied the value for manure application should be zero (0). In case the information on a particular practice is not provided the data should be "not available".

Apart from the metadata there is other data that is essential to correct interpretation of the result of the trial. For example, the yield response to treatments cannot be interpreted correctly without some information on the rainfall, to determine at least the sufficiency of the rainfall. For the all tables that contain yield data the 'year', 'crop name' and the 'number of observations' are the mandatory fields. The 'year' is part of the identifier of the Trial and therefore needs to be entered. The 'crop name' is part of the identifier for the yield data record; it needs to be entered such that it is known to which crop the yield data refers to. In the description of the experiment the names of the test crops are also specified, but an experiment can be conducted on several crops at the same time, like in an intercropping system, while the yield data is provided for that specific, individual crop. The 'number of observations' is required because in the meta-analyses this parameter is used to determine whether there is any significant difference between the results obtained from the various experiments (or it determined the weight of the contribution of this specific experiment to the evidence of the treatment effect generated from the various experiments combined). Of course, yield data itself and some statistical measure of the variance is assumed to be entered.

The data and information that we consider essential for the ISFM legacy database is the data and information that is required if we want to make further use of the date for comparative analyses or, indeed, for synthesis of results from the various experiments. It relates to the 'applicability' requirement of data in 'open access' policies, or the 'fitness for use'. This data and information needs to be provided by the research paper or article, and if not it should not be included in the database. At the same time, it provides guidelines for how to write a good research paper (or at least provide guidelines on what data should be included in the paper) and as such it also provides guidelines for how agronomic trials should be conducted (or at least guidelines for the data to be collected for each experiment or trial to make the data re-usable). Discussions on minimum data sets (or essential data sets) have been held in the past and are still being conducted. Even though it is difficult to reach a conclusion on what these are (because the data that is required depends on the purpose for which is or will be

used), it is useful to provide guidelines on the minimum required data to improve the quality of the data and the relevance of the data (or the ability of the data to be re-used). It is in the interest of the ISFM legacy database that additional data is added in future.

In Appendix 4, we list the proposed parameters of the essential data set for agronomic trials as the outcome of internal brainstorm session at IITA. It defines in broad categories data on which information should be provided, without going into details or being prescriptive on exactly the type of data required. It concurs with the type of data contained in the ISFM legacy database. That is, the database definition provides further details on the descriptors used for describing these variables. For example, descriptors are provided for describing previous land use. Also, it differs in what is considered as essential. For example, 'historical yield data' is not considered essential and the database system does also not cater for entering of data on this subject. The list identifies the data on the following categories as essential, without being prescriptive on the precise descriptors to be used:

- (Geographical) location
- Terrain and site description
- Previous land use
- Historical yield data by crop
- Crop management practices
- Planting material
- Soil properties
- Water availability
- Rainfall

A similar discussion can be held with respect to protocols. Protocols assist in data being collected in a systematic and uniform way and help to improve data quality. However, we should not waste efforts on standardizing protocols, because the protocols will depend on the specific objectives for the field experiment. Rather we will make protocols used for the various agronomic experiment available online, to be used as example for those who want to develop protocols for their own experiment. It is good practice to document the protocols and make these available together with data that is submitted for the data file repository or for entering in the database.

6. Concluding remarks

The activity 1.6a was defined as "to develop a database/repository for ISFM data and information, to populate the database and provide access to the stakeholders. Broadly speaking this activity was about developing and providing facilities for the management of ISFM data. It is a rather broad topic and to develop such a database across institutions and countries is rather ambitious as well. In this project, we have concentrated on developing the ISFM legacy database, to be understood as a database that is developed to capture data from published papers and research documents on past agronomic trials that have a relevance to integrated soil fertility management. So, this is not about the trial data itself; that is the data recorded for each of the individual trials, but rather the processed data and results from these trials. We also consider such an undertaking - to develop a database to capture the data from individual experiments and trials - unrealistic. Rather, the concept is to have each project develop its own database, contribute to the data (file) repositories and extract from those databases or repositories that data that can then be used for some generic analyses to

provide evidence for the effectiveness of ISFM practices. The data from past experiments, certainly from those conducted years ago, are probably of relatively little value or relevance. Nevertheless, it is important to have an overview of what has been done in the past and to capture the relevant data from those experiments and evaluate results to give direction to future research on ISFM within the countries.

We do recognise that there is a need to make more effective use of all the data that is being generated. That requires open access to this data and for this policies for open access need to be adopted and implemented at institutional level especially. As a project, there is little that we can do, apart from advocating for open access policies and laying the ground work for making sharing of data and information possible. We have been advocating for the sharing of ISFM data from agronomic trials through the country soil health consortiums and by working jointly on the ISFM legacy database. At the same time we have been laying the ground work by working on the vocabulary, the metadata standards and design of the database. We need a vocabulary to talk the same language such that we can share information. We need metadata standards of ISFM data for developing the repositories of agronomic trial data files. The data contained in these files needs to be documented and we will use the same metadata standards that is used for documenting the processed data from these trials. Having a generic database structure will help the development of project specific databases and will facilitated the extraction and exchange of data from these databases. So, all these are critical building blocks in developing a data management systems and indeed to facilitate the interoperability and herewith also promoting the shared use of the data.

The ISFM legacy database is currently not available online. We encourage the institutes hosting the CSHC to put at least some of the metadata online such that the discoverability of the data sets is enhanced. The baton is with the research institutions and the country soil health consortiums to develop this further and bring this forward. IITA is developing its data repository. Including data from agronomic trials and these will we open access within one year. This will include data sets from agronomic trials and we will make an effort to have some legacy data still included in that repository.

Appendix 1 Vocabulary of agronomic trials for soil fertility management

Vocabulary	Description
Biomass	The total mass of an organisms referring to the plant or crop in a given area or volume. It usually refers to the biological materials that are not used for food or feed (the harvestable or economic yield). It must be indicated whether the biomass refers to only the above ground plant parts or whether it includes the above and below ground (roots) parts of the plant. This is measured on fresh weight or dry weight basis and expressed in kg/ha or metric tons per ha.
Biomass fresh weight	Fresh weight of the biomass refers to the weight of the biomass when just harvested; that is before any loss of moisture has occurred due to drying.
Biomass dry weight	Weight of the dry biomass, obtained by drying of the biomass harvested, or calculated based on the moisture percentage determined based on drying a sample of the biomass harvested. The drying is done in an oven or in the air.
Block	A group of experimental units, i.e. treatments. The 'block' has significance in the context of statistical design of field experiments, in which the treatments are replicated. The grouping of the treatments (blocking) is done to control the variation (assuming that the conditions in the group/block is more homogenous than between blocks) and for purpose of efficiency. We call the design complete block design if the complete set of treatments is replicated in blocks. In case not the complete set of treatments is replicated it is referred to as an incomplete block design.
Cob	In maize, the cob refers to the central core of an ear of maize on which the kernels grow. The ear is also considered a cob, but is not fully a "cob" until the ear is shucked, or removed from the plant material surrounding the ear. Synonym or related terms: Ear
Cover crop	A cover crop is a crop planted to cover the soil with the primary aim to prevent or limit soil erosion. Sometimes there are secondary objectives, like in case of legume cover crops to improve soil fertility or to provide fodder and feed for livestock. Cover crops may have benefits in terms of better control of pest and diseases, control of weeds, improved soil biodiversity and other Synonyms or related terms: Green Manure, Improved fallow.
Crop	A crop is a cultivated plant that is grown as food, feed, fruit or vegetable. All crops are cultivars, produced through breeding and selection. The common name for a crop often refers to the plant species or subspecies,

	whereas the specific cultivar or variety relates to the taxonomic level below that of plants species or sub-species, the cultivar rank in taxonomy.
Crop residue	Crop residue refers to an organic material usually the remains of crops left over after the economic part of the crop has been harvested. This includes leaves, stalks (stems), husks, seedpods and nonedible roots. Crop residue can be used for many purposes: e.g. to manage soil organic matter (improve soil quality, to cover the soil), used for animal feed, for fuel, building material and other.
Crop residue management	Refers to how crop residue is managed: crop residue is removed from the field, left on the soil surface, burned or incorporated into the soil.
Cultivar	A plant variety that has been produced from a natural species and is maintained by cultivation. Cultivar are generally produced through selective breeding. Synonym or related terms: Variety
Ear	The seed-bearing head or spike of a cereal plant.
Ear leaf	The leaf attached directly below an ear of maize
Experiment	In the present context an experiment refers to the one or more agronomic trials conducted in the field, replicated over different locations and in time over different seasons, following scientific procedures, and adopting a particular experimental design, to find evidence for the efficacy of certain products or management practices in improving different aspects of crop production under varying conditions, to investigate the effect of environmental conditions on the response of crops to the use of the products or application of management practices or to demonstrate effectiveness of agronomic practices on the crop production. The experiment is identified by the experimental feature: what is it that is exactly being studied or demonstrated? An article may report on the different aspects of an experiment or study, in which case it is good to separate these and treat these as separate experiments. This often occurs when multiple factors are considered but when the data does not allow to evaluate the effect of each combinations of treatment factors individually.
Experimental feature	The experimental feature indicates what exactly is studied or investigated. This generally refers to the effects of a particular treatment factor on the outcome variable. In case of multifactorial designs this generally refers to the interaction effect of the treatment factors and on the outcome variable, which in our context generally refers to yield.
Factor	A circumstance, fact or influence that contributes to a result. In an agronomic context and experimental setting this refers in general to the effect or influence of the factor on crop production. The factor refers to the treatment factor. Experiments can be differentiated based on the number of factors that are considered. In a multi-factorial experiment or design, the interactions

	between the factors are investigated through the combination of treatments. In a single factor experiment only one factor varies while all other conditions remain constant. Synonym or related terms: Variable
Fertilizer	Fertilizer is any material organic or inorganic applied to the soil, on foliar parts of the plant or as seed coating to supply nutrients for plant growth and development. Without further specification, the term fertilizer is generally used to mean chemical or inorganic fertilizer.
Flowering	The time to flowering is often measured in trials as the time when 50% flowering has been reached. In maize, this refers to when the plant starts to tassel, shed pollen, or extrude silks.
Fertilizer application rate	This refers to the quantity of a fertilizer of a particular type applied per unit of area, expressed in Kg per ha,
Field	The field refers to a defined piece of land that is used for planting of crops; in the context of agronomic trials to that piece of land where a particular trial is conducted. When trials are conducted on farmer's fields, the field refers to the particular piece of land on the farm used for planting of crops, often with fixed or semi-permanent boundaries. Fields (both on-farm and on-station) have a land use and cropping history, as well as a management history and have soil and topographic characteristics that will be of influence on the outcome of the trial. Synonym or related terms: Location, site
Grain	**Grains** are small, hard, dry seeds, with or without attached hulls or fruit layers, harvested for human or animal consumption. Agronomists also call the **plants** producing such seeds "**grain crops**". The two main types of commercial **grain crops** are cereals such as **wheat** and rye, and legumes such as beans and soybeans
Grain yield	The grain yield refers to amount of grain harvested from cereals or grain legumes, and is used to distinguish from other parts of the plant being harvested and from other type of crops (e.g. root and tuber yield). The grain yield is generally expressed in dry matter to make comparison possible.
Green manure	Green manures are fast-growing plants sown to cover bare soil. Often used in the vegetable garden, their foliage smothers weeds and their roots prevent soil erosion. When dug into the ground while still green, they return valuable nutrients to the soil and improve soil structure. Green manures are often leguminous plants, which can take up nitrogen from the air, and as such provides for a 'free' source of nitrogen.
Head	The grain sorghum head is a panicle with spikelets in pairs. For sorghum, it is often referred to as 'head' rather than the 'ear' that is used for other cereal crops, like wheat or maize.
Improved fallow	Improved fallow is defined as the targeted use of planted species in order to achieve one or more of the aims of fallowing within a shorter time or on a smaller area compared to natural fallow. Improved fallow

	can be seen as a promising method to increase the productivity of smallholder farming systems in the tropics by reducing the fallowing period needed to restore soil fertility. Synonyms or related terms: Cover crop, green manure.
Inorganic fertilizer	Any fertilizer of inorganic or chemical nature, used to supply crop nutrients. It generally refers to the compounds that are the result of some chemical process, which does not necessarily imply an industrial process. Rock phosphate, for example, is also considered an inorganic fertilizer (though it may originate from organic materials) Synonym or related terms: chemical fertilizer
Integrated Soil Fertility Management	A set of soil fertility management practices that necessarily include the use of fertilizer, organic inputs and improved germplasm, combined with the knowledge on how to adapt these practices to local conditions, aiming at optimizing agronomic use efficiency of the applied nutrients and improving crop productivity. All inputs need to be managed following sound agronomic and economic principles
Intercropping system	This refers to the practice in which two or more crops are planted in the same field, often in different arrangements of rows of the different crops interlaced or alternated. The most common goal is to get greater yield by making more efficient use of the land and water resources.
Inter-row spacing	If planting is done in rows this value represents the distance between he rows.
Intra-row spacing	If planting is done in rows this value represents the distance between the planting stations or planting hills within the row.
Kernel	Is the synonym for grain in a maize plant
Least Significant Difference (LSD)	The LSD calculates the smallest significant difference between two means, as if a test has been run on those two means (as opposed to all of the groups together). When the LSD indicates a significant difference this means that at least one of the groups tested differs from the other groups. .
Manure	Manure is an organic material that is used to fertilize land, usually consisting of animal excreta (from e.g. poultry, cattle, goat, etc.) with or without accompanying litter such as straw, hay, or bedding materials, collected and dumped in a pit or heaped and allowed to rot and decompose to various degrees. When domestic waste like ashes etc. are added, it is referred to as 'farm yard manure' (FYM). Distinction is made with 'green manure' when the origin is from plant material only and when applied to the soil as fresh material, and 'compost' which is well rotted especially plant material.
Mulch	A mulch is a layer of material applied to the surface of an area of soil. Its purpose is any or all of the following: ❖ to conserve moisture ❖ to improve the fertility and health of the soil

	❖ to reduce weed growth ❖ to enhance the visual appeal of the area A mulch is usually, but not exclusively, organic in nature. Organic materials used for mulching may refer to grass clippings, leaves or straw, wood chips, saw dust, or shredded bark. It may be permanent (e.g. plastic sheeting) or temporary (e.g. bark chips). It may be applied to bare soil, or around existing plants. Mulches of organic nature will be incorporated naturally into the soil by the activity of worms, ants and other organisms. The process is used both in crop production and in gardening, and when applied correctly can dramatically improve soil productivity. The distinction with use of crop residue is found in that in case of mulch the material is brought from outside the field for the specific purpose of mulching.
Multi-locational trials	Trials of the same design conducted in multiple locations within a given area and within a given season that are all part of the same study aiming to get insight in the spatial variability in crop performance and response to particular treatments.
Net plot	Net plot refers to that part of the plot that is used for harvesting of the crop in a particular trial or from which samples (soil and plant) are taken to determine soil and crop characteristics of that particular plot. Generally the outer rows and the outer plants at the beginning and end of the row are not harvested to discount for possible border influences. The plants harvested from the net plot are considered to give a better and more accurate estimate of the yield or other variables being measured and therefore response to the treatment. The size of the net plot is generally specified in the protocol for the trial
Nutrient application	The nutrient application rate generally specified in Kg/ha. The form or compound to which the application rate refers should always be provided. Preferred is the nutrient expressed in elementary form (e.g. P instead of P2O5).
Nutrient omission trial	This refers to trials in which nutrient limitations or deficiencies in the soil are determined by leaving out one of nutrients in subsequent treatments from the complement of nutrients that are applied at rates that are considered non-limiting. There should be one treatment with all the nutrient applied for reference.
Nutrient response trial	This is referred to all trials conducted with the aim to evaluate the responsiveness of a particular crop to a specific nutrient input, which may refer to the input of macro-nutrients, secondary nutrients or micro-nutrients. Nutrients are applied in often different application rates, as single nutrient or in combination with other nutrients, in order to determine the response curve and optimum application rate.
Organic Fertilizer	A fertilizer of organic nature, derived from animal matter, animal or human excreta or vegetable matter (e.g. compost, and crop residue) aimed at supplying crop nutrients to the soil or directly to the plant.

	Organic fertilizers, by adding organic material to the soil have added benefits to just the supply of nutrients, which may even be the prime reason for applying organic fertilizer Synonym or related terms: Manure
Organic resource management	We refer here to organic resources that are used for the management of soil organic matter and soil fertility. In the present context, we use the term organic resource management to denote a management practice in which plant materials are directly applied onto the land. The plant materials may refer to cuttings of Calliandra, Tephrosia, Tithonia or other. This is to distinguish this practice from practices like the use of compost, green manure and use of animal manure that are also classified as organic resources.
Plant density	Number of individual plants per ha. It can be (automatically) calculated if the plant spacing (Inter-row and intra-row) is known, together with the number of plant per planting station.
Planting date	This refers to the time in which crops are planted or sown within a year.
Plot	A plot is the smallest spatial unit in a trial; that is the part of field that is allocated for a particular treatment. The plot relates to the Treatment like the Field relates to a Trial. There is only one treatment on a plot at a particular time. In this system, we do not allow for subplot as a separate entity. In a split-plot design, one plot with a particular treatment in one season is split in two parts in the following season, each with their own specific treatment and are therefore considered as separate plots. This is modelled through defining the relationship between the instances of this entity, the plots from different trials conducted in different seasons. Information on the split plot design is given in the description of the experiment (experimental design).
Replicate	The replication or repetition of the same treatment on several plots or experimental units. within a trial. It is done to achieve consistency in the results of the experiment. One particular treatment within a trial may be replicated (e.g. the control treatment), several of the treatments or the complete set of treatments in trial may be replicated. If a treatment is truly effective, the averaging effect of replication will reflect the potential benefit of a treatment. Replication reduces variability in experimental results, and increases the significance and the confidence level from which a researcher can draw conclusions.
Season	In agriculture, the word “season” refers to the growing season, the part of the year during which the local weather conditions favour plant growth. Each crop has a specific growing season, and the season may therefore vary from one crop to the other, varying in start and end date and in duration of the growing period and time required for the crop to reach maturity. Time of planting and time of harvesting may vary between crops and even between crop varieties, it is generally possible

	to group the growing seasons for broad categories of crops for a particular agro-ecological zone. The starting date of the season in Nigeria for Cassava, for example, coincides with the start of the rain season (April/May) but ends in March the following year (Although the rain season ends in October/November) In our database the season as such is not indicated, but rather derived from the location and the time the crop was planted. In this way, it can also be determined whether several crops are grown consecutively within the same season.
Site	A 'site' may have different connotations. It may refer to a particular location where a trial or an experiment is conducted, In which case it is referred to as field in our vocabulary. It may also refer to a larger area where several trials are conducted. Project may use different definitions and use different names for the different types of sites. In our system we have not adopted a particular definition for site and it is not identified as a specific entity.
Soil depth	The soil depth or effective soil depth is the depth of the soil to which the roots of a plant can readily penetrate, generally the depth of a barrier layer, where the barrier layer can be rock, a cemented or compacted layer (e.g. iron pan, hard pans/plough pan) or can be assocaited with a sudden transition in soil texture (from clay to sand or from sandy of loamy to clay, for example)
Soil fertility	The ability of the soil to supply essential plant nutrients and soil water in adequate amounts and proportions for plant growth and reproduction in the absence of toxic substances which may inhibit plant growth (FAO).
Soil health	Soil health is the capacity of soil to function as a living system, with ecosystem and land use boundaries, to sustain plant and animal productivity, maintain or enhance water and air quality, and promote plant and animal health.
Soil Organic Matter (SOM) trial	Soil organic matter trial is a trial that involves any material produced originally by living organisms (plant or animal) that is returned to the soil and goes through the decomposition process. In a SOM trial different organic resources are generally tested with respect to effects on soil quality and soil health. http://www.fao.org/docrep/009/a0100e/a0100e04.htm
Standard Error (SE)	The standard error of the mean (SE of the mean) estimates the variability between sample means that you would obtain if you took multiple samples from the same population. It is calculated by the dividing the standard deviation of the sample, divided by the square root of the number of observations (N)
Stem diameter	This refers to the measurement of the plant stem above the ground (either as circumference from which the diameter can be derived or as measurement of the stem diameter in two perpendicular directions).

	The stem diameter together with plant height are often sued as indicators of plant growth.
Stover	Maize stover consists of the leaves and stalks of maize (Zea mays ssp. mays L.) plants left in a field after harvest and consists of the residue: stalk; the leaf, husk, and cob remaining in the field following the harvest of cereal grain. Stover makes up about half of the yield of a crop and is similar to straw. In trials the stover weight is often measured in the field. This will generally exclude the cob weight.
Test crop	This is the crop that is used in a trial for measuring growth and yield parameters. In an intercropping system, the test crop refers to the crop that is considered most important. In a maize-based system the maize is defined as the test crop and the crop used for intercropping is called the 'intercrop'
Tillage	Tillage is the agricultural preparation of soil by mechanical agitation of the soil such as by digging, stirring, and overturning. This can be done using hand tools, by using animal traction of mechanized traction. Synonyms or related terms: Land preparation (though land preparation may include activities like land clearing, weeding etc.)
Treatment	Controlled application of a product or practice in an experimental trial such as: seeding rate, fertilizer rate or fertilizer type, insecticide or fungicide application timing or rate, crop varieties, etc. The treatments are tested to determine impact on crop growth, yield and/or quality of a particular product and/or practice. In our system, the treatment refers to the complete set of agronomic practices and/or products that apply to that treatment and the description of the treatment should therefore also include the description of all these management practices and products that are applied. This concerns the land, soil, crop, nutrient and water management practices that are applied. No distinction is made between those practices that vary between the different treatments in the trial and those that remain constant or the same. Commonly the treatment is denoted by only those variables (or factors) that vary across the treatments in the experiment, but the description nevertheless needs to contain the full information on the agronomic practices applied. These practices might vary between different, but similar, experiments and needs to be fully documented to allow for comparison of the results from these various experiments. It will also allow to better interpret the results of the trial. The number of treatments in an experiment/or trial depends on the factors (or variables) considered and the number of variations allowed for each of these factors. In a single-factor design the number of treatments is equal to the number of variations for this factor (e.g. different fertilizer applications rates). In a multi-factor experiment the potential number of treatments

	is equal the product of the number of variation of each of the factors considered.
Trial	An agronomic trial is a particular instance of an experiment that is conducted in one particular locations (field) at a particular point in time (season), to test the effect of particular conditions and/or management practices (and their possible interactions) on the performance, behavior or quality of a crop. This definition equates a trial with field trial and excludes the experiments that are conducted in the laboratory or in a green house, like pot trials. Such type of trials requires a different structure for their description and storage of the data and therefore different database design. The trial design is determined by the experimental setup; that is by a specific set of treatments and a statistical design of the experiment. This implies that if any of the treatments change or the design changes, the experiment changes and you have a different trial. A trial can be replicated at different locations or in different seasons, but it then will become a separate, individual trial.
Trial design	Statistical design of the experiment/trial that determines the layout of the experimental units or plots. The experiment can consider one or more factors in its design. The following designs can be considered: • Single Block or non-replicated design - strip design • Complete Randomized Design • Randomized Complete Block Design (Blocks of equal size, each containing all treatments) • Latin Square Design • Incomplete Block Design • Group Balanced Block Design
Trial management	This refers to the different types of trials as defined by the various roles of the researcher and farmer in the design, management and implementation of the trial. The FAO distinguishes between these trial types: 1. Researcher managed and researcher implemented (RMRI) 2. Researcher managed and farmer implemented (RMFI) 3. Farmer managed and implemented (FMFI) (http://www.fao.org/docrep/v5330e/V5330e0b.htm#9.2 Types of trials)
Trial type	The type of trial depends on the purpose of the experiment. The following trial types are defined: ❖ Variety trial (variety selection) ❖ Provenance trial (testing of suitability of provenances – from different geographical areas) ❖ Fertilizer trials: ○ Nutrient response trial (to determine response to different nutrient application rates)

	○ Nutrient omission trial (to determine nutrient limitation in the soil) ❖ Agronomic/cultural trials ○ Tillage practice (to determine the effect of the various tillage and land preparation practices) ○ Soil Organic Matter trial (to test the effect of various soil organic amendments) ❖ Chemical (other than fertilizer) trial
Variety	The crop variety refers to the cultivar rank in taxonomy, which is the rank below species or sub-species. The name is often derived from a naming scheme which consists of letters and numbers (like "TME 419") Synonym or related terms: Cultivar
Yield	Yield refers to the harvestable component of a crop.

Appendix 2. Metadata set for ISFM legacy data

Table 3 Administrative metadata - information related to the source of the data

	Administrative metadata / Data related to the source of the data	
Attribute/ descriptor	**Description**	**Remark/relationship**
Experiment identifier	Unique code and identifier of the experiment or study of which the results are presented in the article or for which the data is provided.	This can be a custom code or name, not strictly part of the metadata set, but rather part of the ISFM legacy database
Citation	Full citation of the publication that reports on the data set (data publication) or of the scientific publication that reports the results of the study or experiment (referring to results from field trials); Includes authors, year of publication, title, journal or publisher	The publication is considered the data source. If there is no publication (thesis, article or report) this field should not be filled.
Creator	Name of the creator or curator of the data; name of the person from whom the data can be sourced; can be the person who oversaw the experiment or trials	The scientist who was responsible for the experiment or trials is often the only source of the data or of further information on the data or trials if required.
Date created	Year when the data set was created, which may be different from the publication date	Refers to the trials data; which is not the publication date, not the date when the experiment or trials were conducted
Resource location (URL)	URL where the publication can be obtained or is sourced from	For the ISFM legacy database this refers to the URL where the publication can be sourced.
Data source location (URL)	URL where the trial data can be sourced from	This is the location where the supporting data for the article or publication can be sourced from; May refer to processed and as well as the unprocessed ('raw') data
Project name	Name of the project under which the data was generated	Use the acronym under which the project is known or use full title with specification of the acronym between brackets)
Institution	Lead institute of the project	
Contact person	Name of the contact person at the lead institute (or from the project) through whom the data or information on the data can be obtained	
Contact email	Email address on the contact person	

Restriction in use	Indicate whether the data is open access, or whether there are any restrictions in the use of the data. And if there are restriction what kind of restrictions	This applies to the data that is not already available through the publication

Table 4 Descriptive metadata - data describing the experiment or trial

	Descriptive metadata/ data describing the experiment or trials	
Attribute/ descriptor	**Description**	**Remarks**
Experiment component identifier	Code or identifier for the component of the experiment or study; that is for the data set related to the particular aspect of the research	This can be a letter or number. Together with the identifier of the experiment it provides a unique identifier for the data related to this specific aspect or component of the study or experiment. Is not strictly part of metadata
Type of trial	Indicate the type of trial as indicated in the value list; The type of trial often explains the objective of the experiment or trial (e.g. variety selection, identifying nutrient limitations, determining response to fertilizer application, etc.)	See value list below. For as yet we are only considering data obtained from trials, but we are aiming to extend the scope to also include agronomic survey data for example.
Experimental factors	Name the experimental factor(s) in the experiment/trial; that is related to those variables that are considered in the treatments (e.g. N application rate, plant density, crop variety and other);	
Experimental feature	Explain what was determined, what was studied; generally, the effect of a factor, or combination of factors, on crop performance (e.g. relation or effect of combined use of inorganic and organic fertilizer on maize crop yield, or. effect of use of manure on the agronomic efficiency of fertilizer application on poor soils).	Can be derived from the expected outcomes of the trial or experiment and is related to the objectives of the study. The experimental feature should accurately reflect the data that is provided in the article or data set rather than what is described in the text (e.g. if different weeding regimes are part of the treatment layout, but results are not presented for these different weeding regimes, it should not be included in the feature statement)
Type of experiment	Indicates whether the trials are done on-station or on-farm	For the ISMF legacy database we are only considering these two possibilities. But in principle this descriptor could also include pot experiments in a greenhouse for example
Trial Management	Describes how the trial was managed (see value list). Indicates whether the	Only three values are possible: Researcher managed and

	trial/experiment is designed and managed by the scientist or the farmer.	implemented, research managed but implemented by the farmer, and farmer managed and implemented
First or primary test crop	Provides the name of the main crop used for testing in the experiment/trials	This name should always be provided because we do not have trials without a test crop.
Secondary crop	Name of the secondary crop in case of intercropping system	If the name of a second crop is given this implies that the crops are intercropped. The first crop is always one of the 'major' crops (e.g. in a maize-legume system maize will be considered the first or primary crop and the legume the secondary crop)
Tertiary crop	Name of the third crop in the intercropping system	What is the second or third crop in an intercropping system is to some extend arbitrary, but should be based on the importance of the crop (from economic perspective or from food security perspective),
Relay crop	Name of the crop used in the relay cropping systems	
Start date	The year the experiment started and first year for which data was collected and presented	
End date	The year the experiment was closed and the last year for which data was collected and included in the data set of presented in the publication	Difference between end and start date determines the study duration expressed in years
Number of seasons	Specifies the number of seasons the trials were conducted and for which data was collected and presented (irrespective this is for a continuous period or intermittent)	Irrespective of whether there are one or two seasons in a year. If a trial is irrigated and conducted outside the season it should also be counted.
Experiment location	Indicate the area or location where the experiment is conducted. That is the area to which the locations for the agronomic trials are confined. Here you specify the administrative unit at lowest administrative level that still encompasses the experimental area.	This maybe be the village or community, the district, province, state or other, whatever applies to the country concerned. The specific location of the individual trials is specified elsewhere.
Country	Specify the country where the trials are conducted	This implies that the data is organized per country, which only in exceptional cases will not be the case
Agro-ecological zone (AEZ)	The agro ecological zone the area belongs to where the data was collected.	Allowed values are specified by a value list.
Data content	Specifies whether the data set contains data on soils, on crop parameters like yield and crop growth data and on the weather (especially rainfall)	

Table 5 Descriptive metadata - technical information to help interpretation of the data

	Descriptive metadata / technical information	
Attributes	**Description**	**Remarks / relations**
Sampling strategy	Sampling strategy for the selection of sites or locations for the agronomic trials that are part of the experiment or study.	Indicates how the locations for the trials are selected, even if it concerns one trials and location only.
Trial design	Statistical design of the experiment/trial that determines the layout of the experimental units or plots. Specifies the number of factors (one, two or three) with the statistical design adopted as per value list.	See value list for the possible statistical designs for agronomic trials
Data collection method	Describes the tools, techniques or methods used in data collection. This refers to the collections of soil data, crop and plant data and weather data.	For soils, how soil samples were taken and analysed (what methods or tools were used); similar for collections of plant growth, crop and yield data (how samples are taken, measurement techniques and analytical methods), and similar for rainfall and other weather data: how was the data obtained. Information is generally provided in the articles or otherwise can be obtained from the protocol.

Table 6 Descriptive metadata - information about the treatments as part of the trial design

Attribute / descriptor	**Description**	**Notes**
Treatment name	Name or code given to the treatment	Arbitrary name or code that can be the code given to identify the treatment in the data source
Variety or Genotype	Specifies the specific variety or cultivar of the primary test crop used in this treatment.	In a variety trial the treatment will have its specific variety being tested. In trials where different varieties are not tested the name of the variety is the same for each treatment. The name can be either the scientific name or the common name if the common name is well recognized and uniquely identifies the variety.
Variety 2nd crop	The name of the variety of the crop used as first intercrop	Allows to specify the name of the specific variety of the intercrop used in the experiment, but also allows to specify the variety of the intercrop used in specific treatments, though that is seldom investigated
Variety 3rd crop	The name of the variety of the crop used as second intercrop	

Plant density test crop	The plant density in number of plants per ha of the test crop or major crop	The plant density is provided irrespective of whether this crop is used as single crop or whether it is being intercropped
Plant density first intercrop	The plant density in number of plants per ha of the first intercrop	
Plant density second intercrop	The plant density in number of plants per ha of the second intercrop	
Plant density relay crop	plant density in number of plants per ha for the crop used as relay crop	
Cover crop	Name of the cover crop used in a fallow rotation as green manure in the previous season	The 'cover crop' is synonym for 'Improved fallow' or 'green manure', to cover the soil and improve soil fertility and refers to the crop grown in the season preceding the season for which the data is presented
Nitrogen application	Application rate of N in Kg ha-1 applied by means of mineral fertilizer in the treatment	Total N supplied by means of the application of mineral (or inorganic) fertilizer(s); should not include the N applied through other means, like manure
Phosphorous application	Application rate of P expressed in Kg of elemental P per ha for the treatment	To get P application by converting from P2O5 application rate, multiply by 0.437; does not include the P applied through means other than chemical fertilizer (e.g. from manures or other)
Potassium application	K application rate expressed in Kg elemental K per ha for the treatment	To get K application rate from K2O application rate, multiply by 0.83
Boron application	Boron application rate in Kg per ha in the treatment	Application rate to be specified as B (elementary B) and requires conversion if application rate is specified for Boron trioxide, boric acid or other.
Copper application	Copper application in Kg Cu per ha for the treatment	
Zinc application	Zinc application in Kg of elemental Zn per ha for this treatment	Conversion is required if Zn application is given as zinc sulphate (contains 36% Zn) or zinc oxide (contains 78% Zn), which are probably the most common form in which zinc is applied
Calcium application	Calcium application in Kg per ha for this treatment	
Magnesium application	Magnesium application expressed in Kg Mg per ha for this treatment	
Manganese application	Manganese application expressed in Kg elemental Mn per ha	Requires conversion if Mn application is specified as MgO (41-68 % Mn), MnO_2 (Mn is 63%), Manganese Sulphate (26-28% Mn) or other.

Sulphur application	Sulphur application in Kg elemental S per ha for this treatment	
Fertiliser placement method	Method of fertiliser application for the starter or basal application as per value list	Refers to the basal fertilizer application; Controlled by value list; Fertilizer is either placed by broadcasting (surface or incorporated), applied as concentrated strips (bands, on the rows or between the rows, surface applied or deep placed – 8 -15cm), localized (spot applied under or to the side of the seed) or applied together with the seed (seed placement or 'pop-up'); other application methods are with irrigation water (fertigation) and as foliar application
Type of organic manure	Type of organic manure by origin or source of the manure and/or compost according to value list	Controlled by value list – For compost the source material can be listed or added to 'compost'; organo-mineral fertilizers are to be treated as both application of manure and as application of the mineral fertilizer (the application rates for the various elements need to be entered) and the organic component needs to be recorded under this attribute
Manure application rate	Application rate of the manure in Kg/ha	
Manure placement	Placement method of the organic manure	In general, this is either broadcast (or spread out) or applied per pocket as often the case in the Zaï system
Bio-fertilizers	Type of bio-fertilizer used in the treatment	Controlled by value list
Lime application rate	Application rate of agricultural lime in Kg/ha	
Crop residue management	Specify the method of crop residue management as per value list	Specifies whether crop residues are left on the field or taken off, and when left on the field whether it is burned, left on the surface or incorporated
Mulching operation	Specifies the material used as mulch as per value list	Materials for mulching can vary from synthetic materials and plastics to pebbles, straw or grass clippings. The materials are brought into the field from outside, to distinguish between the use of crop residue as management practice
Tillage operation	Type of tillage operation applied on the plot for this treatment	Provide value as specified in the value list. In case of animal traction you can specify the type of animal used in addition. In case of mechanical tillage, you can add whether it is done by two-wheel or four-wheel machinery. Also the type of equipment can be mentioned in addition (.e.g ripper/planter, mouldboard, disc, chisel)

Water supply method	Means by which water is supplied to the crop is managed according to value list provided	Water supply is either classified as rainfed, post-flooding (making use of residual water in the soil), or irrigated. Information can be added on how water supply/ infiltration is improved through water harvesting techniques, like tied ridging and use of planting pits (zaï), half-moon etc.). For irrigated systems, the type of irrigation can be added
Pest and Disease management	Type of practice for the control of pest and diseases, according to value list	Can be either preventive measures and natural control, biological control mechanisms or chemical control
Weed management practice	Type of practice to control weeds	Is either 'no weeding', 'hand weeding', mechanical weeding' or 'chemical weed control'; the frequency of weeding can be specified if that is part of the treatment factors and the activity ingredient can be specified in case of the chemical weed control, together with the application rate in case that is a treatment factor
Organic resources ''cut & carry'	Organic resource used, by source and resource quality	Refers to other organic resources than manure or compost, and crop residues, used for soil fertility improvement and that is carried from outside the field.

Table 7 Structural metadata - information assisting in how to get and use the data

	Structural metadata / Information on how the data is organized and structured	
Missing data code	Specify the code used for missing data in the data set	For example, 'NULL' for data that is missing and 'NA' for 'not applicable'
Data set structure	Indicate whether the data set is part of a series, or larger data set and specifies how the data sets (files) are organized and structured.	Research may be conducted in multiple countries and covering different type of trials. It is then important to specify how the data is organized and for example how that is reflected in the convention for naming the files. This only applies to data sets and repositories and not the research papers
Data integrity	Indicates whether the data can be relied on and procedures followed to assure the quality of the data. That is whether the data is checked for completeness, consistency, accuracy and reliability (errors in the data) on logical integrity.	Papers may provide information on the data quality or procedures followed for quality assurance. For data sets and databases there may be separate documentation on the quality assurance protocols. For example, how the data has been checked for systematic and random errors (e.g. outliers), on consistency (e.g. whether the same units have been applied for all entries) and whether the internal links and relationships are properly defined and implemented.
File format	Specifies the format of the file by which the data is stored and/or delivered	This applies to research articles and reports in what format the are available (e.g. WORD or PDF) as well as to data files the format in which

		the data is stored (e.g. comma separated file (csv), EXCELL file, or other)
Data type	Specify in which form the data is presented: text, table, graph	This applies especially to the research articles and scientific publications, the data is either presented as textual, tabular or graphical data.

Table 8 Data level metadata - structure of the data dictionary and information facilitating use of the data

	Data level metadata	
Attributes	**Description**	**Remarks**
Attribute name	Full name of the attribute	
Attribute label	Label used in the database/data set to denote the attribute	
Description	Description of the attribute	
Data type	Whether text or number	In some databases, you will have to specify the data format (e.g. maximum number of characters if text, or precision)
Unit	Specification of the unit in which the data is described	
Class/code	The classification system that applies to the specific attribute	If the data refers to the class value the classification codes and classification system needs to be explained
Derived algorithm	Specification of the algorithms through which the attribute value is determined.	If the data is derived from other data through calculation, the algorithm needs to be specified
Weighing variable	Specifying the weight of the variable If the attribute is used in a calculation that uses weighing factors, e.g. weighted average	
Value list	Specify the value list of allowed attribute values	If the entry of data is controlled by a list of allowed attribute values, specify the domain

Appendix 3 Value list for the various metadata descriptors

Table 9 Value lists for the descriptors of the administrative and descriptive metadata

Descriptor/attribute	Value list
Type of trial	Variety trial
	Provenance trial
	Nutrient Omission Trial (fertilizer)
	Nutrient Response Trial (fertilizer)
	Soil Amendment Trial – (organic matter, lime and other)
	Cultural management trials (pre-planting practices)
	Cultural management trials (planting and post planting practices)
	Other - Unspecified
Type of experiment (field experiment)	On station-single location – short duration
	Long term trial (on-station - single location – long duration)-
	On farm – single location
	On farm- multiple location
Trial management	RMRI – Researcher managed and implemented
	RMFI – Researcher managed and farmer implemented
	FMFI – Farmer managed and farmer implemented
Trial design	Single block or strip design – non-replicated design
	CRD – Complete randomized design
	RCBD – Randomized complete block design
	Latin Square design
	Incomplete Block design
	GBBD – Group balanced block design
Agro-Ecological Zone	Sahelian
	Sudanian savanna
	Northern Guinea savanna
	Southern Guinea savanna
	Derived savanna
	Mid-altitude savanna
	Rainforest
	Mangrove swamps

Table 2. Value list for the treatment descriptors

Descriptor/Attribute	Value list
Fertilizer placement methods	Broadcast – surface
	Broadcast – incorporated
	Band – row – incorporated
	Band – on row – surface

	Band – inter-row Localized (spot) – dibble Localized (spot) – mechanical Seed placement Deep placement Injection Fertigation Foliar application Not applicable
Bio-fertilizer	Not applicable Rhizobium inoculant Free-living nitrogen fixing bacteria (diazotrophs) Mycorrhiza (AMF) Blue-green algae Azolla Other
Manure (organic)	No organic manure Unspecified Farm Yard Manure (FYM) Poultry manure Cattle manure Pig manure Compost (unspecified) Vermicompost
Manure placement	Broadcast Pocket
Crop residue management	No crop residue on field Stubble Stem and leaves Husks Seed pods Non-edible roots *Specify in combination with treatment (below)* Burned Left on the field Incorporated
Mulching	No mulching Unspecified Gravel or pebbles Grass clippings Leaf and/or straw Wood chips / shredded bark Saw dust Cocoa hulls Pine straw
Tillage operation	No-till Minimum tillage- planting basins Minimum tillage – ripping Strip-till Hand-held hoe Plough – Animal traction

	Mechanical traction (two-wheel/ four wheel) Deep tillage
Water supply I	Rainfed Post-flooding Tied ridges Basins (Zaï, Half moon) Irrigation – surface/gravity Irrigation – sprinkler Irrigation- drip
Pest and Disease management	No P&D management Preventive/ natural control Biological control Chemical control
Weed management	No weeding Manual - once 1 Manual - twice or more Chemical (specify active ingredient and frequency) Mechanical (specify frequency) Biological weed control

Appendix 4 – ISFM legacy database description – Tables

Table 10 The 'field' table - descriptors of the entity 'Field'

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
Identifier	ID	Unique identifier for the trial field	INT	Auto Enter;			
Experiment reference	Exper_ref	Code for the experiment that the trial conducted in this field belongs to	Text	Not null		Defines n:1 relationship between the field and the experiment	
Field referential number	Field_ref	Identifier for the field within the referenced experiment	INT	Not null; default value: 1	Provides together with the Exper_ref the primary key; is added for practical purposes, because the ID alone in principle would be enough		
Location name	Loc_name	Name of the location (village or community) where the field is located	Text	Optional; Can only be edited during data entry			
Farmer Name	Farmer name	Name of farmer in whose farm trial field is placed.	Text	Optional; Can only be edited during data entry			
Field latitude	Field_lat	Latitude of the field.	Double/ Number	Optional; Only edited during the data entry	Decimal degrees, which should be negative (-) for values south of the Equator and positive (+) North of the Equator.		

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
Field longitude	Field_long	longitude of the field	Double/ Number	Optional; Only edited during data entry	Decimal degrees - negative (-) for values west of the Greenwich Meridian and positive (+) East of the Greenwich Meridian.		
Elevation	Field_elev	Elevation of the field	Number	Optional	MASL (meters above sea level)		
Topographic position	Field_topo	Topographic position of the field	Text	Optional; value list			'summit', 'upper slope', 'mid slope', 'foot slope', 'toe slope', 'valley bottom', 'basin floor', 'plain', 'terrace'
Years cultivated	Years_cult	The number of years the field has been cultivated	Number	Optional; Only edited during data entry	Number of years since the land has been cleared for cultivation (since deforestation, land reclamation)		
Type of fallow	Fallow_typ	Type of fallow	Text	Not null; value list	For 'natural fallow' the dominant plant species can be specified; For 'improved fallow' type of crop of plant species can be listed		'no fallow', 'bare land', 'natural fallow', 'improved fallow'
Fallow duration	Fallow_dur	Duration of the fallow period	Number	Optional	Number of seasons; refers to the fallow system in recent years, or otherwise the number of season the field was fallowed in the last 5 years		
Soil classification FAO	Soil_class_FAO	Soil classification according to the FAO classification system	Text	Optional; value list	According to FAO-UNESCO soil map of the world - revised classification systems 1997		

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
Slope_class	Slope_class	Slope class according to USDA system	Text	Optional; by value list			'flat to almost flat', 'gently sloping', 'sloping', 'steep'
Soil depth	Soil _depth	Soil depth class	Text	Optional; value list	Soil depth class as per value list provided		Very shallow (<25cm), Shallow (25-50cm), Moderately deep (50-100cm), Deep (100-150cm), Very deep (>150cm)

Table 11 Soil characteristics table - descriptors for soil characteristics

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
Identifier	ID	Unique identifier for the record of soil characteristics data	INT	Auto Enter;			
Experiment reference	Exper_ref	Code for the experiment	Text	Not null	Soils at different locations can be characterized but all part of the same experiment	Identifies the experiment for which the soil characteristics are determined [n:1 relationship]	
Field referential number	Field_ref	Identifier for the field within the referenced experiment	INT	Not null; default value: 1; use 0 to indicate that the soil characteristics data refers to the summary statistics for the several trial	Arbitrary number assigned to the field to identify the field that may be one of the several trials sites at which the experiment is conducted. Soil characteristics can be	Defines n:1 relationship between the soil characteristics and the field o	

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
				sites used in the experiment	determined for the same field at different times; The soil characteristics data (record) may also refer to the summary statistics for several of the trial site locations.	trial site location	
Year trial establishment	Year_trial	Year in which the trial is established	DATE	Optional	Soil characteristics can be established one or more times for the same trial/field (e.g. before and after); in case the soil characteristics do not refer to one particular trial the reference to the trial is not provided	Defines n:1 relationship to the trial	
Year of sample collection	Year_sample	Year when the samples were collected	date	Not null;	Based on the year of the sample collection it is determined whether the soil characteristics refer to those before or after the trials is conducted.		
Treatment Identifier	Trt_id	Arbitrary number assigned to each treatment in the experiment	INT	Optional; needs to correspond to ID of the Treatment in the Treatment table	Is filled in case the soil characteristics are determined for each plot separately, assuming that within the block of the trial treatments are not replicated (there are no two or more plots with the same treatment	Provides link to the Treatment in the Treatment table	
Soil texture	Soil_text	Soil texture class	Text	Optional; value list	Soil texture class according to USDA soil texture classification system (soil texture triangle), or, if this information is not available,		'Coarse to very coarse', 'moderately coarse',

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
					the soil texture class as determined by the field method.		'medium', 'moderately fine'
Soil Organic Carbon	SOC(%)	Soil Organic Carbon in weight percentage	Decimal	Optional	Percentage		
Total Nitrogen	TN(%)	Total nitrogen concentration in soil	Decimal	Optional	percentage		
pH in water	pH_H2O	pH measured in water	Decimal	Optional	Dilution 1 part soil to 2 parts water		
Available phosphorous	P (mg/kg)	Available P	Decimal	Optional; ppm	Method for determining available P should be recorded		
Aluminium	Al(mg/kg)	Extractable Aluminium from soil in ppm	Decimal	Optional; ppm	Method of extraction should be listed		
Extractable Calcium	Ca (cmol/kg)	Extractable Ca2+	Decimal	Optional	Method of extraction should be listed; conversion required if concentration is provided in different unit		
Extractable Magnesium	Mg (cmol/kg)	Extractable Mg2+	Decimal	Optional	List method of extraction; convert if concentration is provided in different unit		
Extractable Potassium	K (cmol/kg)	Extractable K	Decimal	Optional	Conversion required if concentration is provided in another unit		
Extractable Sodium	Na (cmol/kg)	Extractable Na+	Decimal	Optional	List extraction method and convert the data if provided in a different unit		
Extractable Zinc	Zn (ppm)	Extractable Zn	Integer	Optional	List extraction method and convert data if provided in different unit.		
Extractable Manganese	Mn (ppm)	Extractable Mn	Integer	Optional	List extraction method; convert to correct unit		

Attribute name	Attribute label	Description	Data Type	Validation	Unit / Comments	Relationship	Value lists/Class
Extractable Iron	Fe (ppm)	Extractable Fe	Integer	Optional	List extraction method; convert to correct unit		
Extractable Copper	Cu (ppm)	Extractable Cu	Integer	Optional	List extraction method; convert to correct unit		
Extractable Boron	B (ppm)	Extractable B	Integer	Optional	List extraction method; convert to correct unit		
Cation Exchange Capacity	CEC	Cation Exchange Capacity in cmol/kg	decimal	Optional	Cation exchange capacity of the soil in cmol/kg		
Effective Cation Exchange Capacity	ECEC	Effective Cation Exchange Capacity (cmolc/kg)	decimal	Optional	Total amount of the effective cation exchange capacity of the soil in cmol/kg		
Base saturation	BS %	Base saturation in %	decimal	Optional	Percentage of fraction of exchange cations of CEC		
Bulk density	BD	Bulk density in g/cm3	decimal	Optional	The weight of the soil in given volume		
Total porosity	Por_tot	Total porosity in %	decimal	Optional	The amount of pore space in a volume of soil in percentage		

Table 12 The 'grain yield' table (cereals and grain legumes)

Attribute name	Attribute label	Description	Data type	Data validation	Comment	Relationship	Value lists
Identifier	ID	Unique identifier for the object/ record	Integer	Auto entry			
Experiment reference	Exper_ref	Code/Identifier for the experiment	Text	Not null		Foreign key; Defines relation with the experiment in the Experiment/data source table	

Field reference number	Field_ref	Arbitrary number or code to identify the field	TEXT	Not null;	Use “00” to indicate if results are obtained from several trials conducted at various locations	Foreign key; Defines the relation with the field described in the Field table and soil characteristics
Year	Year_trial	Year when the trial was implemented	date	Not null; enter ‘9999’ for if the data is averaged for the various trials conducted at the specified location		Foreign key; Exper_ref + Field_ref + Year together provides the relationship to the trial.
Treatment Identifier	Trt_id	Arbitrary number assigned to each treatment in the experiment	INT	Not-null;	Exper_ref + Trial_ref (Field_Ref+Year) + Trt_Id provides unique identifier for the Treatment	Foreign key; Provides link to the Treatment description in the Treatment table.
Crop Name	Crop_name	Name of the crop for which the yield data is provided	Varchar	Not null;	Is used distinguish between the different crops if data for more than one crop is provided a part of the same experiment	
Number of Observations	No_observ	Number of observations for deriving the statistical data	Integer	Not null;	Refers to the number of replications, determined by the number of replications within the trial times the number of trials over which the data is averaged	
Grain yield fresh weight	Yld_grain_fw	Mean of the grain yield in fresh weight	Kg/ha	numeric		
STD grain yield fresh weight	Yld_grain_fw _SD	Standard deviation of fresh	Kg/ha	numeric		

		weight of grain yield		
SE grain yield fresh weight	Yld_grain_fw_SE	Standard error of the mean fresh weight of grain yield	Kg/ha	numeric
LSD grain yield fresh weight	Yld_grain_fw_LSD	LSD of the mean grain yield in fresh weight	Kg/ha	numeric
Grain yield dry weight	Yld_grain_dw	Mean of grain yield dry weight dry	Kg/ha	numeric
STD grain yield dry weight	Yld_grain_dw_SD	Standard deviation of grain yield dry weight	Kg/ha	numeric
SE grain yield dry weight	Yld_grain_dw_SE	Standard error of the mean dry weight of grain yield	Kg/ha	numeric
LSD grain yield dry weight	Yld_grain_dw_LSD	Least significant difference of grain yield dry weight	Kg/ha	numeric
Stover yield fresh weight	Yld_stover_fw	Mean of the stover yield fresh weight	Kg/ha	numeric
STD stover yield fresh weight	Yld_stover_fw_STD	Standard deviation of stover yield fresh weight	Kg/ha	numeric
SE stover yield fresh weight of stover yield	Yld_stover_fw_SE	Standard error of fresh weight of stover yield	Kg/ha	numeric

LSD of fresh weight of stover yield	Yld_stover_fw_LSD	LSD of fresh weight of stover yield	Kg/ha	numeric	
Dry weight stover yield	Yld_stover_dw	Mean of stover yield dry weight	Kg/ha	numeric	
STD stover yield dry weight	Yld_stover_dw_STD	Standard deviation of dry weight of stover yield	Kg/ha	numeric	
SE stover yield dry weight	Yld_stover_dw_SE	Standard error of stover yield dry weight	Kg/ha	numeric	
LSD of dry weight of stover yield	Yld_stover_dw_LSD	LSD of dry weight of stover yield	Kg/ha	numeric	
Grain100/1000 dry weight	Grn100/1000-dw	Mean weight of 100 (or 1000) dry grains	Kg/ha	numeric	100 grains for large sized grains (e.g. maize); 1000 grains for small grains like sorghum
SD of 100/1000 grain dry weight	Grn100_dw_SD	Standard deviation of 100/1000 grains dry weight	g	numeric	100 grains for large sized grains (e.g. maize); 1000 grains for small grains like sorghum
SE 100/1000 grains dry weight	Grn100_dw_SE	Standard error of 100/1000 seed dry weight	g	numeric	100 grains for large sized grains (e.g. maize); 1000 grains for small grains like sorghum
LSD 100/1000 grains dry weight	Grn100_dw_LSD	Least significant difference of 100 (or 1000) grains dry weight seed dry weight	Kg/ha	numeric	For large grains like maize the weight of 100 grains is measured; in case of small grains, like of sorghum, it is the weight of 1000 grains

Table 13 The 'root and tuber yield' table

Attribute name	Attribute label	Description	Data type	Data validation	Comment	Relationship	Value lists
Identifier	ID	Unique identifier for the object/ record	Integer	Auto entry			
Experiment reference	Exper_ref	Code/Identifier for the experiment	Text	Not null		Foreign key; Defines relation with the experiment in the Experiment/data source table	
Field reference number	Field_ref	Arbitrary number or code to identify the field	TEXT	Not null;	Use “99” to indicate if results are obtained from several trials conducted at various locations	Foreign key; Combined with the experiment ID defines the relation with the field described in the Field table	
Year trial establishment	Year_trial	Year when the trial was implemented/established	date	Not null; enter ‘9999’ if the data does not refer to one particular trial	Exper_ref + Field_ref + Year_trial together identifies the trial; time of harvesting after planting is specified in the treatment table	Part of the foreign key; defines relationship with the ‘trial’ data	
Treatment Identifier	Trt_id	Arbitrary number assigned to each treatment in the experiment	Text	Not-null; needs to correspond to ID of the Treatment in the Treatment Table	Exper_ref + Trial_ref (Field_Ref+Year_trial) + Trt_Id provides unique identifier for the Treatment, irrespective of how may time the treatment is replicated within the trial	Foreign key; Provides link to the Treatment description in the Treatment table	
Crop Name	Crop_name	Name of the crop for which the yield data is provided	Varchar	Not null;	Is used to distinguish between the different root & tuber crops if the experiment		

					includes more than one tuber or root crop.
Number of Observations	No_observ	Number of observations over which the values are averaged	Integer	Not null;	Refer to the number of replications of the treatment in the trial, times the number of blocks, times the number of trials over which the data is averaged
Tuber yield fresh weight	Yld_tuber_fw	Average Yield of total fresh tuber in t/ha	Decimal	Optional	t/ha; relates to all tubers harvested independent of size
Standard deviation tuber fresh weight yield	Yld_tuber_fw_SD	Standard deviation of the seed weight in t/ha	Decimal; t/ha	Optional	t/ha
Standard error tuber yield fresh weight	Yld_tuber_fw_SE	Standard error of the mean of the seed yield	Decimal; t/ha	Optional; if SD and N is given the SE can be calculated	Is only entered if the standard deviation is not given
Least significant difference tuber yield fresh weight	Yld_tuber_fw_LSD	Least significant difference of shoot fresh weight yield in t/ha	Decimal; t/ha	Optional	t/ha
Tuber yield dry weight	Yld_tuber_dw	Average Tuber yield expressed in dry weight in t/ha-1	Decimal;	Optional	t/ha; Method on how dry weight is determined need to be specified in the metadata.
Standard deviation tuber dry weight yield	Yld_tuber_dw_SD	Standard deviation of the seed weight in t/ha	Decimal;	Optional	t/ha;
Standard error tuber yield dry weight	Yld_tuber_dw_SE	Standard error of the mean of the seed yield	Decimal;	Optional	t/ha;

Least significant difference tuber yield dry weight	Yld_tuber_dw_LSD	Least significant difference of shoot fresh weight yield in t/ha	Decimal; t/ha	Optional	t/ha;
Tuber Number	No_tubers	Average number of tubers per plant	Decimal	Optional	
Standard deviation number of tubers	No_tubers_SD	Standard deviation of number of tubers per plant	Decimal	Optional	
Standard error of number of tubers	No_tubers_SE	Standard error of the mean of number of tubers per plant	Decimal	Optional	
Least significant difference number tubers	No_tubers_LSD	LSD in number of tubers per plant	Decimal	Optional	
Weight percentage commercial sized tubers	Yld_tuber_comm	Average yield percentage of commercial sized tubers from total yield	Decimal	Optional;	Indicates the percentage of the total yield that relates to commercial sized tubers; if the figure provided relates to commercial sized tubers only the percentage should be 100%
Number of commercial sized tubers per plant	No_tuber_comm	Average number of commercial sized tubers per plant	Decimal	Optional	Applies to tubers and roots alike; if the previous number of roots already refers to the commercial sized root only the same number should be provided

Table 14 The 'banana yield' table

Attribute name	Attribute label	Description	Data type	Data validation	Comment/Unit	Relationship	Value lists
Identifier	Id	Unique identifier for the banana yield record	Text	Auto entry			
Experiment reference	Exper_ref	Code/Identifier for the experiment	Text	Not null		Foreign key; Defines relation with the experiment in the Experiment description and data source table	
Field reference number	Field_ref	Arbitrary number or code to identify the field	TEXT	Not null;	Use “99” to indicate if results are obtained from several trials conducted at various locations	Foreign key; Defines, together with ‘Exper_ref’ the relation with the field described in the Field table	
Start year of the trial	Year_trial	Year when the trial was established	date	Not null;	Enter ‘9999’ if the results are summarized over the years the trial is conducted	Foreign key; ‘Exper_ref’ + ‘Field_ref’ + ‘Year_trial’ establishes the relation to the trial	
Treatment Identifier	Trt_id	Arbitrary number assigned to each treatment in the experiment	Text	Not-null; needs to correspond to ID of the Treatment in the Treatment table	Exper_ref + Trial_ref (Field_Ref+Year_trial) + Trt_Id provides unique identifier for the Treatment	Foreign key; Provides link to the Treatment in the Treatment table	
Crop name	CropName	The name of the crop, the yield measurement was taken	Varcha	Allow editing only during data entry; Indexing none			

Attribute name	Attribute label	Description	Data type	Data validation	Comment/Unit	Relationship	Value lists
Cycle number	Cycle_no	The cycle number for which the data is provided	Text	Not null; enter "0" when the data is averaged over two or more cycles	Bunches are harvested from the same mat in cycles; data is provided for each cycle separately or summed or averaged for several subsequent cycles		
Number of Observations	No_observ	Number of observations for the calculation of the mean, STD or SE, referring to the number of bunches, or average number of bunches per cycle	Integer	Not null;	Refers to the number of bunches for that particular cycle for which the data is collected, or to the total number of bunches, summed for the various cycles considered.		
Bunch weight	Bunch_weight	Average bunch weight	Decimal	Not null; Allow editing only during data entry. Indexing none	Kg		
Standard deviation bunch weight	SD_Bunch_wt	Standard deviation from the mean of the bunch weight	Decimal	Optional	Kg		
Standard error bunch weight	SE_Bunch_wt	Standard error of the mean bunch weight	Decimal	Optional	Kg		
Least Significant difference bunch weight	LSD_Bunch_wt	Least significant different between groups of mean bunch weight	Decimal	Optional	Kg		
Number of hands	No_hands	Mean of the number of hands per bunch	Decimal	Optional; indexing none			

Attribute name	Attribute label	Description	Data type	Data validation	Comment/Unit	Relationship	Value lists
Standard deviation number of hands	SD_no_hands	Standard deviation of the number of hands per bunch	Decimal	Optional			
Standard error number of hands	SE_no_hands	Standard error of the mean of the number of hands	Decimal	Optional			
Least Significant difference number of hands	LSD_no_hands	Least significant different between groups of mean of the number of hands	Decimal	Optional			
Number of fingers	No_fingers	Average number of fingers per bunch	Decimal	Optional			
Standard deviation number of fingers per bunch	SD_no_fingers	Standard deviation from the mean of the bunch weight	Decimal	Optional			
Standard error mean number of fingers	SE_no_fingers	Standard error of the mean bunch weight	Decimal	Optional			
Least Significant Difference number of fingers	LSD_no_fingers	Least significant different between groups of mean bunch weight	Decimal	Optional			

Table 15 The vegetable yield table

Attribute name	Attribute label	Description	Data type	Data validation	Comment	Relationship	Value lists
Identifier	ID	Unique identifier for the record	Integer	Auto entry			
Experiment reference	Exper_ref	Code/Identifier for the experiment	Text	Not null		Foreign key; Defines relation with the experiment in the Experiment description / data source table	
Field reference number	Field_ref	Arbitrary number or code to identify the field	TEXT	Not null; entered from the Field table	Enter "99" if data is averaged for two or more locations where trials are conducted	Foreign key; Defines the relation with the field described in the Field table	
Year of trial establishment	Year_trial	Year when the trial was established	date	Not null;	Enter "9999" when the data is averaged over trials conducted in two or more seasons.	Foreign key; Exper_ref + Field_ref + Year_trial provides the link to the trial	
Treatment Identifier	Trt_id	Arbitrary number assigned to each treatment in the experiment	Text	Not-null; needs to correspond to ID of the Treatment in the Treatment table		Foreign key; Exper_ref + Trial_ref (Field_Ref+Year) + Trt_Id provides link to the Treatment in the Treatment table	
Crop name	Crop_name	Name of the crop for which the yield data is provided	Varchar	Not null;	Is used distinguish between the different crops if data for more than one vegetable crop		

Attribute name	Attribute label	Description	Data type	Data validation	Comment	Relationship	Value lists
					is provided a part of the same experiment		
Number of Observations	No_observ	Number of observations used in the calculation of the statistics	Integer	Not null;	Refers to the number of treatment replications in the trial, times the number of blocks, times the number of trials over which the data is averaged		
Yield shoot fresh weight	Yld_shoot_fw	Average of fresh shoot weight harvested in Kg/ha	Decimal; Kg/ha	Optional;	Kg/ha		
Standard deviation shoot fresh weight yield	Yld_shoot_fw_SD	Standard deviation of the shoot fresh weight in Kg/ha	Decimal; Kg/ha	Optional	Kg/ha		
Standard error shoot fresh weight yield	Yld_shoot_fw_SE	Standard error of the mean of the shoot fresh weight yield	Decimal; Kg/ha	Optional; if SD and N is given the SE can be calculated	Kg/ha; Is only entered if the standard deviation is not given		
Least significant difference shoot fresh weight yield	Yld_shoot_fw_LSD	Least significant difference of shoot fresh weight yield in Kg/ha	Decimal;	Optional			
Yield shoot dry weight	Yld_shoot_dw	Dry weight yield of shoot in Kg/ha	Decimal	Optional			
Standard deviation shoot dry weight yield	Yld_shoot_dw_SD	Standard deviation of the shoot fresh weight in Kg/ha	Decimal; Kg/ha	Optional			
Standard error shoot dry weight yield	Yld_shoot_dw_SE	Standard error of the mean of the	Decimal; Kg/ha	Optional; if SD and N is given the SE can be calculated	Is only entered if the standard deviation is not given		

Attribute name	Attribute label	Description	Data type	Data validation	Comment	Relationship	Value lists
		shoot fresh weight yield					
Least significant difference shoot dry weight yield	Yld_shoot_dw_LSD	Least significant difference of shoot fresh weight yield in Kg/ha	Decimal;	Optional			
Fruit Yield	Yld_fruit	Average of the fruit yield in Kg/ha	Number; Kg/ha	Optional	Kg/ha; Fruit in the scientific sense which also includes nuts		
Standard deviation fruit yield	Yld_fruits_SD	Standard deviation of the shoot fresh weight	Decimal; Kg/ha	Optional	Kg/ha		
Standard error fruit yield	Yld_fruits_SE	Standard error of the mean of the shoot fresh weight yield	Decimal; Kg/ha	Optional; if SD and N is given the SE can be calculated	Kg/ha; Is only entered if the standard deviation is not given		
Least significant difference fruit yield	Yld_fruits_LSD	Least significant difference of shoot fresh weight yield	Decimal;	Optional	Kg/ha;		
Seed weight	Yld_seed	Mean of the seed yield	Decimal	Optional	Kg/ha		
Standard deviation of the shoot dry weight yield	Yld_seed_SD	Standard deviation of the seed weight in Kg/ha	Decimal; Kg/ha	Optional	Kg/ha		
Standard error seed yield	Yld_seed_SE	Standard error of the mean of the seed yield	Decimal; Kg/ha	Optional; if SD and N is given the SE can be calculated	Kg/ha; Is only entered if the standard deviation is not given		
Least significant difference seed yield	Yld_seed_LSD	Least significant difference of shoot fresh weight yield	Decimal;	Optional	Kg/ha		

The water supply table

Attribute name	Attribute label	Description	Data type	Data validation	Comment	Relationship	Value lists
Identifier	ID	Unique identifier for the record	Integer	Auto entry			
Experiment reference	Exper_ref	Code/Identifier for the experiment	Text	Not null		Foreign key; Defines relation with the experiment in the Experiment description / data source table	
Field reference number	Field_ref	Arbitrary number or code to identify the field	TEXT	Not null; entered from the Field table	Enter "99" if rainfall data is not provided for a specific location but rather for the area where the trials are conducted	Foreign key; Defines the relation with the field described in the Field table	
Year of trial establishment	Year_trial	Year when the trial was established	date	Not null;	Enter "9999" when the data is averaged over trials conducted in two or more seasons, or whether they are long terms averages.	Foreign key; Exper_ref + Field_ref + Year_trial provides the link to the trial	
Cumulative water supply	CWS	Cumulative water supply at harvest	integer	Optional	mm; refers to cumulative rainfall for rainfed systems with the amount of water supplied for irrigated systems; may refer to data recorded at location or projected from data recorded at nearby stations or from general rainfall statistics for the area		

Appendix 5. Proposed parameters of the essential data set for agronomic trials

Results of the discussion conducted internally at IITA to identify the minimum set of data to be collected at each site for which agronomic trials are conducted. The idea was to either present these as guidelines or to make these compulsory for all agronomic trials conducted by IITA. The parameters below are considered the be essential to facilitate the proper interpretation of the data and results from the trial and for evaluation of the outcome from field trials.

- **Farm location:** lowest level of administrative unit (state, municipality, province)
- **Geographical coordinates** (arrange to have all GPSs coordinates in the same format, or provide tools/algorithms for the conversion from one format to the other)
- **Terrain** (short description of the field site: slope (slope length and steepness), inclination (abrupt changes in slope), position within the landscape - such as plateau, slope, base of hill/mountain, valley bottom, etc. -, description of the near surrounding land use and land cover / field borders - such as bordering forest, river or swamp with description of the dominant border vegetation -, outcrops and any other feature (tree stumps, depressions termite hills within the field site, outside the used site, etc.)
- **Previous land use**: crop rotation or previous crop(s) grown in the same field or fallow length and type (grass, bush, forest, planted legumes etc.) in previous years. (We need to have descriptors/classes for land use description, to include something on cropping index / land use intensity parameters
- **Historical yield data by crop** (period to be determined and depending on fallow phases.) (Nice if you can have it, but not sure whether it is essential. Depends on what you are going to do with the information.)
- **Farmer's current crop management practices:** tillage, burning, crop residue management, nutrient inputs, herbicide and pesticide use, any other crop management operation).
- **Planting material:** source, variety, maturity or growth duration, any other crop specific traits – Let's make a distinction between planting materials as part of the management practices – or for the specific trial. The latter seems obvious, the information on the planting material used by the farmer is only relevant if the yield on farmer's fields is going to be surveyed and the information is going to be used to compare with the results from the trials. We need separate table to describe the trial design.

- **Soil chemical properties:** pH, OC, TN, exch. K, Ca, Mg, avail. P and texture – (this will require a concrete protocol on the number of samples to be taken per unit area, the minimum sampling depth and the depth increments for sampling - for instance minimum sampling depth is 0-20cm to be taken in 0-10 and 10-20 cm increments or in increments that are compatible with the 0-10 and 10-20cm such as 0-5, 5-10, but not 0-12 or 0-7.5cm. To have guidelines on the number of samples per unit area we would need to do some statistical analyses determine how many samples are required to have repeatable soil chemistry data – the point at which the variance will not further increase. This could as well be depending on the trial type and specifics of the treatments and their response to differences in soil chemistry. Needs to include protocols for coding of samples
- **Water availability:** {fully rain fed, rain fed with supplemental irrigation, irrigation}; may not be a priority but if horticultural areas are considered it may be important. Maybe to include in site description?
- **Rainfall:** (rain gauge at field site or reference gauge at known distance) – these would be simple manual gauges that we could be bulk purchased. Training of staff or farmers to read these gauges may be required. Right so this should be part of guidelines for the data collection on the trial, rather than the site.

The Ghana Soil Health Consortium