v.2 February 2010

International Union of Soil Sciences
PROPOSAL FOR A WORKING GROUP ON SOIL INFORMATION STANDARDS (WG-SIS)

The Working Group would operate under the auspices of the Commissions on Soil Geography (C1.2) and Pedometrics (C1.5)

Prepared on behalf of Commissions 1.2 and 1.5 by

Peter Wilson, CSIRO, Canberra, Australia (peter.wilson@csiro.au)

with contributions from

Alex McBratney (The University of Sydney, Sydney, Australia)
Alex.McBratney@usyd.edu.au
Sonya Ahamed (CIESIN EI Columbia, USA)
sahamed@ciesin.columbia.edu
David Jacquier (CSIRO, Canberra, Australia)
david.jacquier@csiro.au
Jim Fortner (NRCS, USA)
jim.fortner@lin.usda.gov
Rob Lokers (WUR, Wageningen, The Netherlands)
Rob.Lokers@wur.nl
Amir Pourabdollah (University of Nottingham, UK)
axp@Cs.Nott.AC.UK
Johan Leenaars (ISRIC World Soil Information, Wageningen, The Netherlands)
johan.leenaars@wur.nl
Bob MacMillan (ISRIC World Soil Information, Wageningen, The Netherlands)
bob.macmillan@wur.nl
Alfred Hartemink (ISRIC World Soil Information, Wageningen, The Netherlands)
alfred.hartemink@wur.nl
Luca Montanarella (European Commission - DG JRC, Ispra, Italy)
luca.montanarella@jrc.ec.europa.eu
Simon Cox (European Commission - DG JRC, Ispra, Italy)
simon.cox@jrc.ec.europa.eu

EXECUTIVE SUMMARY

There is an increasing need to collect, collate and share soil data and information within countries, across regions and globally. A number of national and international activities and projects are currently dealing with the issues associated with collation of disparate data sets. Standards are being developed for data storage, transfer and collation such as, for example, in the *GobalSoilMap.net* project, e-SOTER and the EU Inspire GS-SOIL. Individually these will not provide a single internationally recognised and adopted standard for soil data and information exchange.

An opportunity therefore exists for the IUSS to provide leadership and focus in this area through the inauguration of a formal Working Group on Soil Information Standards. This proposal is based on the successful Pedometrics and Digital Soil Mapping Working Group models. The Working Group on Soil Information Standards will provide a mechanism for cross-initiative collaboration and coordination leading to development of an International Standard for Soil Information. It will increase the accessibility and use of soil information for cross sectoral issues. The Working Group would give the SoilML information model both scientific credibility and international standing and strengthen the role of the IUSS. A number of meetings and workshops are being planned to progress the draft SoilML information model.

1. Background to the case

There is an increasing need to collect, collate and share soil data and information within countries, across regions and globally. Timely access to consistent and authoritative data and information is critical to issues related to food production, climate change, water management, energy production and biodiversity. Soil data and information are managed by numerous agencies and organisations using a plethora of processes, scales and standards. A number of national and international activities and projects are currently dealing with the issues associated with collation of disparate data sets. Standards are being developed for data storage, transfer and collation such as, for example, in the *GobalSoilMap.net* project, e-SOTER and the EU Inspire GS-SOIL. Individually these will not provide a single internationally recognised and adopted standard for soil data and information exchange.

A recent *GlobalSoilMap.net* meeting held in Wageningen, The Netherlands, discussed the needs of a harmonized information model for collation of a global 90 metre grid of key soil attributes (organic carbon, soil texture, pH, depth to bedrock/impeding layer, and predictions of bulk density and available water capacity) at six specified depth increments. The meeting considered a number of existing data base implementations (such as ASRIS, NASIS, WISE, SOTER) as well as emerging abstract information models that are being expressed in Unified Modelling Language (UML) (such as e-SOTERML). It examined related information models, such as GeoSciML (see https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/WebHome) and the lessons learnt in developing and implementing such community agreed models, features and vocabularies.

There is a need to develop a global soil information standard, to be called SoilML, that would allow access and use of data across a broad range of international initiatives (such as GEOSS and INSPIRE) as well as supporting national, regional and local data interoperability and integration. The meeting agreed to adopt the interoperability approaches of formalising the information model in UML with XML encoding for data transfer as well as re-using existing features and patterns where appropriate such as those found in GeoSciML and Observations and Measurements (see http://www.opengeospatial.org/standards/om).

2. Goals and Performance Indicators

The goal is to provide collaboration and coordination of soil information standards efforts worldwide by:-

(1) Developing an abstract UML information model for soil data called SoilML (in consideration of existing information standards such as ISO, Observations & Measurements, GML, GeoSciML etc) to be promoted as an International Standard for soil data transfer and collation.
(2) Considering and promoting the role of soil information and standards in GEOSS the Global Earth Observation System of Systems.
(3) Developing an international soil sample numbering system to avoid data translation and duplication errors
(4) Holding annual meetings or workshops in conjunction with IUSS Commissions and national soil science societies.
(5) Providing training workshops in Soil Information Standards to provide capacity building (in conjunction with academic institutions and national and international agencies).

The specific objectives of the Working Group are to:

- to develop a conceptual model of soil information drawing on existing data models
- to implement an agreed subset of this model in an agreed schema language (UML)
- to implement an XML/GML encoding of the model subset
- to develop a test-bed to illustrate the potential of the data model for interchange
- to identify areas that require standardised classifications in order to enable interchange.

Key performance indicators for the success of the Working Group (2010-2015)

(a) One workshop/meeting every year
(b) Increasing interest and membership in the Working Group
(c) Publication of papers and conference proceedings in the field of soil informatics
(d) Recognition and adoption of international standards for soil information
(e) Use of soil information standards by projects in developed and developing countries.
(f) Demonstrate the ability to collate consistent data sets for key soil attributes for significant proportions of the world by 2015.

3. Prior activities

Over the last few years, there has been a rapid growth in interest in this topic worldwide. A number of activities and initiatives are individually developing harmonized approaches to managing soil data and information. Some of these include -

3.1 eSOTER - http://www.esoter.net/

The e-SOTER project started on September 1st 2008 and will last until February 2012. The kick-off meeting was held at ISRIC, Wageningen, from 11 to 12 September, 2008.

WP6 Development of an e-SOTER dissemination platform

WP leader: DG Joint Research Centre, Vit Penizek

Will develop a data dissemination portal for e-SOTER based on the INSPIRE Directive and applying GEOSS Data Sharing Principles. Links with existing or emerging international soil platforms will be explored.
Activities comprise:

1. analysis of data specification and exchange rules (XML). We will prepare the SoTerML, closely linked with GeoSciML.

3.2 Harmonized World Soil Database

http://www.fao.org/nr/lman/abst/lman_080701_en.htm

The Land Use Change and Agriculture Program of IIASA (LUC) and the Food and Agriculture Organization of the United Nations (FAO) have developed a new comprehensive Harmonized World Soil Database (HWSD). The work was carried out in partnership with:

- ISRIC-World Soil Information, together with FAO, were responsible for the development of regional soil and terrain databases and the WISE soil profile database;
- the European Soil Bureau Network, which had recently completed a major update of soil information for Europe and northern Eurasia, and
- the Institute of Soil Science, Chinese Academy of Sciences, which provided the recent 1:1,000,000 scale Soil Map of China.

The HWSD is a 30 arc-second raster database with over 16000 different soil mapping units that combines existing regional and national updates of soil information worldwide (SOTER, ESD, Soil Map of China, WISE) with the information contained within the 1:5,000,000 scale FAO-UNESCO Soil Map of the World (FAO, 19711981).

The resulting raster database consists of 21600 rows and 43200 columns, which are linked to harmonized soil property data. The use of a standardized structure allows for the linkage of the attribute data with the raster map to display or query the composition in terms of soil units and the characterization of selected soil parameters (organic Carbon, pH, water storage capacity, soil depth, cation exchange capacity of the soil and the clay fraction, total exchangeable nutrients, lime and gypsum contents, sodium exchange percentage, salinity, textural class and granulometry).

Further expansion and update of the HWSD is foreseen for the near future, notably with the excellent databases held in the USA: Natural Resources Conservation Service US General Soil Map (STATSGO) http://www.ncgc.nrcs.usda.gov/products/datasets/statsgo, Canada: Agriculture and AgriFood Canada: The National Soil Database (NSDB) http://sis.agr.gc.ca/cansis/nsdb and Australia: CSIRO, ACLEP, Natural Heritage Trust and National Land and Water Resources Audit: Australian Soil Resource Information System ASRIS http://www.asris.csiro.au/index_other.html.

On the basis of soil parameters provided by HWSD seven key soil qualities important for crop production have been derived, namely: nutrient availability, nutrient retention capacity, rooting conditions, oxygen availability to roots, excess salts, toxicities, and workability.

3.3 GS SOIL *- http://www.gssoil.eu/*

Assessment and strategic development of EU INSPIRE compliant Geodata-Services for European Soil Data. Initial meeting held June 2009 Hanover, Germany.
Seven working packages are defined - including: WP2 Content Provision Framework; WP3 Data Management and Meta-Data; WP4 Harmonisation and Semantic Interoperability.
Recently completed a review of requirements and context existing for harmonised soil datasets at a scale larger than 1:1million - to be extended through online "Stakeholder Questionnaire on Soil Data Requirements" (www.vupop.sk/form/) open until 11th Feb 2010.

GS SOIL registered as Spatial Data Interest Community (SDIC) on the INSPIRE website (http://inspire.jrc.ec.europa.eu)

3.4 GlobalSoilMap.net *- http://www.globalsoilmap.net/*

A consortium that aims to make a new digital soil map of the world using state-of-the-art and emerging technologies for soil mapping and predicting soil properties at fine resolution. This new global soil map will be supplemented by interpretation and functionality options that aim to assist better decisions in a range of global issues like food production and hunger eradication, climate change, and environmental degradation. This is an initiative of the Digital Soil Mapping Working Group of the International Union of Soil Sciences IUSS.

A four day Data Model workshop (Minutes attached at Appendix A) was held in Wageningen, The Netherlands, from 1st to 4th December 2009 with the following objectives:

- To begin the process of defining a conceptual model for storing soil data for a global soil information system (or exchange mechanism) (to be referred to as SoilML)
- To define the scope of *GlobalSoilMap.net* needs to store and distribute spatial data

- To identify, in particular, the scope (range) of soil profile and soil map legacy data that will be supported by the *GlobalSoilMap.net* data model
- To identify and review implementations of existing soil profile or soil map data models to be supported within the *GlobalSoilMap.net* project
- To familiarise a core group of soil scientists and data modellers with current best practices in conceptual information modelling
- To review implementations of existing soil profile/soil map data models using the selected analytical tool and to interpret results
- To identify the next steps to take in the standards setting process.

4. Need for such a Working Group

There is an increasing need for soil data to be brought together in a timely and efficient manner to support national, regional and international issues. The development and adoption of information standards for soil data is therefore critical. A number of activities are currently developing soil information standards for their own use to fulfil the needs of specific project outcomes. Individually these projects will not develop a widely applicable international soil data standard and may not have longer-term, enduring, non-project responsibilities.

Development of an internationally accepted and widely applicable soil data standard requires a strong governance arrangement to ensure coordination and consolidation of efforts. Development will be strengthened, and hopefully hastened, by a formal group in which everyone can participate. The IUSS is ideally placed, as an international scientific committee, to provide the required non-project specific governance umbrella for development and future management of soil information standards. An IUSS working group will provide international scientific credibility as well as a focus for the soils community and others dependant on soil data and information.

This scientific committee working group proposal follows a similar approach adopted by the international Union of Geological Sciences (IUGS) in the development and governance for GeoSciML
(see https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/WebHome).
The Commission for the Management and Application of Geoscience Information (CGI) aims to enable the global exchange of knowledge about geoscience information and systems. This scientific committee approach is preferred as it provides a strong connection to the scientific discipline over a purely standards based approach, such as has been recently adopted in the development of WaterML to be governed through the Open Geospatial Consortium (OGC).

5. Linkage to other disciplines

This working group will develop strong links to other IUSS working groups, particularly pedometrics and digital soil mapping, to ensure the data and information needs of soil science disciplines are catered for. It will also develop linkages to other sectors and disciplines developing similar and associated information models, such as those in the water (WaterML) and geosciences (GeoSciML) arenas. Similarly, strong links will be developed to the information standards community, including adoption and use of existing and developing approaches and standards such as ISO, UML, Observations & Measurements, W3C etc.).

6. Brief plans

Without wishing to pre-empt the deliberations of the Working Group once established, it is probably useful to set out what appear to be the principal issues to be addressed. Some very tentative plans for meetings are also suggested.

Major activities will include

- Developing an abstract UML information model for soil data called SoilML
- Promoting the role of soil data in GEOSS
- Developing an international unique soil sample numbering system
- Meetings and workshop including training in soil information standards

Potential Meetings

2009	Workshop on data standards - GlobalSoilMap.net. Wageningen, The Netherlands December 2009
2010	Workshop to progress the SoilML information model - GlobalSoilMap.net Oceania Node. Canberra, Australia March 2010
2010	Workshop to progress SoilML - Ispra, Italy in connection with the Digital Soil Mapping conference Rome, May 2010
2010	Workshop to progress SoilML and initial trial implementations - Brisbane, Australia in connection with the IUSS World Congress on Soil Science August 2010
2011	Pedometrics 2011?

7. Benefits to the IUSS

An IUSS working group with the purpose of developing, implementing and managing an international soil information standard will provide recognition of soil science as a critical global discipline and enhance the use and accessibility of soil data in many contemporary and future international issues. It will enhance the reputation of the IUSS for guiding and enabling strategically important science.

8. Constitution

The Working Group will be established within the Division D1: Soils in Space and Time, under the Commissions of C1.2: Soil Geography and C1.5 Pedometrics. Officers of the Working Group will be appointed according to IUSS statutes < www.iuss.org >. The contributors to this proposal recommend Mr Peter Wilson (CSIRO, Australia) be appointed Chairperson for the initial term of the Working Group.

Some interim arrangements may have to be made, since it is hoped that the Working Group can become operational in order to quickly progress development of soil data standards and promote the activities of the Working Group at the European Geosciences Union General Assembly in May 2010. Also, an initial meeting of the Working Group is being planned to coincide with GlobalSoilMap meetings and the Digital Soil Map conference to be held in Rome in late May 2010.

Appendix 1
Minutes of the *GlobalSoilMap.net* Data Model Workshop, 1-4th December 2009, Wageningen, The Netherlands.

Report 5
Data model workshop
1-4th December 2009 Wageningen, Netherlands

Present

Alex McBratney (The University of Sydney, Sydney, Australia)
Alex.McBratney@usyd.edu.au
Sonya Ahamed (CIESIN EI Columbia, USA)
sahamed@ciesin.columbia.edu
Peter Wilson (CSIRO, Canberra, Australia)
Peter.wilson@csiro.au
David Jacquier (CSIRO, Canberra, Australia)
david.jacquier@csiro.au
Jim Fortner (NRCS, USA)
jim.fortner@lin.usda.gov
Rob Lokers (WUR, Wageningen, The Netherlands)
Rob.Lokers@wur.nl
Willem van Deursen (Consultant, The Netherlands)
wvandeursen@carthago.nl
Amir Pourabdollah (University of Nottingham, UK)
axp@Cs.Nott.AC.UK
Johan Leenaars (ISRIC - World Soil Information, Wageningen, The Netherlands)
johan.leenaars@wur.nl
Bob MacMillan (ISRIC - World Soil Information, Wageningen, The Netherlands)
bob.macmillan@wur.nl
Alfred Hartemink (ISRIC - World Soil Information, Wageningen, The Netherlands)
alfred.hartemink@wur.nl

Summary

A four-day Data Model workshop was held in Wageningen, The Netherlands, with the following objectives:

- To begin the process of defining a conceptual model for storing soil data for a global soil information system (or exchange mechanism) (to be referred to as SoilML)
- To define the scope of *GlobalSoilMap.net* needs to store and distribute spatial data.
- To identify, in particular, the scope (range) of soil profile and soil map legacy data that will be supported by the *GlobalSoilMap.net* data model.
- To identify and review implementations of existing soil profile or soil map data models we want to support within the GlobalSoilMap.net project.
- To familiarize a core group of soil scientists and data modellers with current best practices in conceptual modelling of database design.
- To review implementations of existing soil profile/soil map data models using the selected analytical tool and to interpret results.

 To identify the next steps to take in the standards setting process.

The following action items were outcomes of the workshop:
 Develop an interim AfSIS data model and entry screen for profile data for use by AfSIS. Agreed upon approach involves also producing a master data base that is fully compliant with O&M, GML and GeoSciML standards. Delivery by 15th January 2010 and roll-out at Arusha meeting on the 25th January.
 Application and illustration of UML modelling test beds. To be led by Peter Wilson with help from Rob Lokers and contributions from the nodes.
 Preparation and submission of a proposal to IUSS to set up a Soil Information Working Group that will focus on: (i) global soil information standards (ISO, O&M, GML, GEO, GeoSciML), (ii) the role of soil information in GEOSS, (iii) the development of an international soil sample numbering system.
 Preparation and distribution of a standard base grid to define locations and unique identifiers for each unique 3 arc-second grid node to be used in the *GlobalSoilMap.net* project. The base grid to be prepared and distributed by Bob by 1st January 2010.
 Selection and establishment of a platform for internal communication, interaction and collaboration for *GlobalSoilMap.net*. Sonya to take the lead. Platform for collaboration and exchange of data and ideas to be up and operational by 15th January, 2010.
Other issues that remain to be addressed: (i) Select and describe a work flow for rescuing and converting-harmonizing soil map legacy data, in particular for Africa. (ii) Identification and acquisition of appropriate cyber-infrastructure for use by the *GlobalSoilMap.net* project. (iii) Development of a plan for how to proceed with specifying the final data model and database design for the *GlobalSoilMap.net* project for delivery of final output products.

1. Agenda

Tuesday 1st December (Chair Alex McBratney)

08:30 – 09:00

Welcome and introductions

09:00 – 10:30

Need for open standards and Lessons from geoscience and hydrology communities (Simon, Peter)

10:30 – 12:30

Presentation of existing candidate data model implementations: ASRIS (David J.),US (Jim),ISRIC (WISE, Soter, SoterML, ISIS) (Johan/Ingrid/Bob)

13:30 – 15:00

Presentation Modeling Approaches Australia (Peter, David, Simon)

15:30– 17:00

Group Discussion on Modeling Approaches

Wednesday 2nd December (Chair Peter Wilson)

09:00 – 09:30

Welcome and review of Tuesday experiences and results (Sonya)

09:30 – 10:30

Presentation of results of UML modeling done prior to workshop (Peter, David, Simon)

10:30 – 12:30

Interpreting Results, Commonalities/ Inconsistencies

13:30 – 14:00

GlobalSoilMap.net Specific Background and Context (Bob)

14:00 – 17:00

Defining the scope of *the GlobalSoilMap.net* model (discussion) Soil site data: **properties;** Soil map data

Thursday 3rd December (Chair Jim Fortner)

09:00 – 09:30

Welcome and review of Wednesday experiences and results (Sonya)

09:30 – 10:30 (Bob)

First steps: Implementing initial procedures for capturing soil profile data for Africa (AfSIS)

10:30 – 17.00

Workshop discussion of how to proceed in developing and implementing an appropriate soil information model

Friday 4th December (Chair David Jacquier)

09:00 – 09:30

Welcome and review of Thursday experiences and results (Sonya)

09:30 – 12:30

Next steps

2. Action Item List

The following list of action items were agreed upon at the workshop.

1. Develop an interim AfSIS data model and entry screen for profile data for use by AfSIS. The agreed upon approach is to use individual `flat file` structures for each unique set of source data sets. Model has fixed fields. The agreed upon approach involves also producing a master data base that is fully compliant with O&M, GML and GeoSciML standards. Procedures will be developed and demonstrated for ingesting data from two or more different `flat files` automatically into the GML compliant master data base. Sonya to assume responsibility for constructing the data model and entry screens in consultation with Bob. Delivery by 15th January 2010 and roll-out at Arusha meeting on the 25th January.
2. Application and illustration of UML modelling test beds. To be led by Peter Wilson with help from Rob and the nodes. A first test bed will be developed by a small group which well then be shared with all the nodes.
3. IUSS Working group - Preparation and submission of a proposal to IUSS to set up a Soil Information Working Group to define a new standard for a global soil information model that conforms to all existing and emerging standards. This proposed Working Group will focus on: (i) global soil information standards (ISO, O&M, GML, GEO, GeoSciML), (ii) the role of soil information in GEOSS, (iii) the development of an international soil sample numbering system. The first draft for the Working Group proposal will be made by Peter Wilson with input from others present at this workshop. Contacts will be made with the IUGS task force to establish standards that are reviewed and approved by an international scientific authority (IUSS - ICSU).
4. Preparation and distribution of a standard base grid to define locations and unique identifiers for each unique 3 arc-second grid node to be used in the *GlobalSoilMap.net* project. Base grid to be prepared by Bob with assistance from Willem and review by Sonya. The base grid to be prepared and distributed by Jan

1, 2010.
5. Selection and establishment of a platform for internal communication, interaction and collaboration for *GlobalSoilMap.net*. Sonya to take the lead in investigating and selecting options, including the Wiki site whose use was demonstrated by Simon Cox. Platform for collaboration and exchange of data and ideas to be up and operational by Jan 15, 2010.
Other issues that remain to be addressed:
1. Select and describe a work flow for rescuing and converting-harmonizing soil map legacy data, in particular for Africa. Bob to develop and circulate a discussion document that outlines ideas and options for a multi-stage process of discovery, rescue, capture and conversion-harmonization of soil legacy map data. Discussion to lead to efforts to define accepted procedures for rescuing, capturing and converting legacy map data to map data that are suitable for use in predicting soil properties for *GlobalSoilMap.net*.
2. Identification and acquisition of appropriate cyber-infrastructure for use by the *GlobalSoilMap.net* project. In the short term there is a need for cyberinfrastructure (hardware, software and human resources) to support the compilation and sharing of data and models for use in the *GlobalSoilMap.net* project proof of concept and operational mapping activities. In the longer term, there is a need to begin to scope out the cyber-infrastructure that will be required to support the eventual delivery of the final products of the *GlobalSoilMap.net* project. Bob and Alfred to engage in a discussion with Sonya and others at CIESIN to develop a plan to collaborate in scoping and designing the cyber-infrastructure needed for the *GlobalSoilMap.net* project.
3. Develop a plan for how to proceed with developing the final data model and database design for the *GlobalSoilMap.net* project for delivery of final output products. Bob to initiate the process of end-user engagement by developing and circulating a discussion document intended to result in identification of a specific person and a strategy for engaging with end users.

3. Summary of discussions

Day 1: Tuesday 1st December (Chair Alex McBratney)

David Jacquier CSIRO

The ASRIS description of issues and problems encountered in identifying and reconciling differences in content and structure between disparate data bases and data models for different states and jurisdictions in Australia was particularly important. David indicated that the final approach was to identify a common subset of attributes and of attribute descriptors (a controlled vocabulary) that all states could agree to and supply. The ASRIS model also required agreement to be reached on how many soils or components were described for each mapped soil entity (map unit) and on the subset of attributes that would be adopted to describe each soil map component.
ASRIS defines a 7 level spatial hierarchy that progresses from level 1 (Division at 30 km) to level 7 (Site at 10 m). The significance of this is that ASRIS found it necessary to define and populate a number of hierarchical spatial layers in which more general data at a coarser resolution is used to establish context for more detailed data at a finer spatial resolution.

There is some dichotomy between ASRIS v1 and v2 and of the technical and institutional issues associated with these two different versions. Conceptually, v1 adopted modelling approaches that principally used point data applied to covariates to predict continuous soil properties and did not make extensive use of existing soil-landscape polygon maps. Several of the contributors to ASRIS expressed concerns with the quality and reliability of the initial v1 maps produced using these methods. The result was production of a second version (v2) that made much more extensive use of existing soil-landscape polygon maps.

Soil property values were computed using area-weighted means of the values of soil properties reported for each soil component in a mapped soil polygon. The importance of recognizing and using existing soil-landscape maps as covariates in the process used to predict continuous soil properties was highlighted. Also highlighted was the importance of maintaining the support and respect of existing soil mapping agencies and communities when applying novel digital soil mapping methods. Ignoring or being insensitive to existing data suppliers and legacy data sets can lead to loss of support for the new prediction methods and for the products that they generate.

Presentation of the 5 layer model of fixed depths for ASRIS profiles led to a discussion of the advantages and disadvantages of fixed depth versus variable depth (horizon) approaches for describing the variation in soil properties with depth in the profile. It was observed that the 5 layer fixed depth model addresses some of the challenges of harmonization and development of standard queries to the data base but that the innovative *GlobalSoilMap.net* approach of using splines to describe continuous variation with depth provides even greater flexibility and more opportunities for interaction and analysis.

A significant aspect of the description of the ASRIS data model was the need that was recognized to adopt different data structures for collation and storage versus presentation and delivery. For collation and storage a (denormalized) data model with few tables, that was easily extensible and easily maintained was favoured. At the core of this data model were three main tables that essentially contained data on features (things), samples (the things that were measured) and results (the values of the things that were measured). For presentation and display, a more flat table structure was required with a single table in which each field contained data for a specific attribute of interest. This structure was optimized for data delivery.

Another significant aspect of the description of the ASRIS project was the importance of achieving agreement on issues related to data ownership and intellectual property (IP) rights. Buy-in by key data providers is essential for success of a collaborative project. Obtaining permission to access and use existing legacy data can be a difficult process that requires sensitivity and persistence. ASRIS found that there was a need for a formal governance structure to encourage data producers to supply their data. The shame factor did not work in getting data producers to supply their data. Only top-level agreement to supply data under agreed conditions was effective in getting data released and shared.

A final aspect of the ASRIS presentation was the issue of reporting measures of uncertainty associated with predictions of soil properties. ASRIS did this using expert

judgment. Alex McBratney emphasised the point that *GlobalSoilMap.net* will be unique in its use of actual measured data to assign measures of uncertainty to each estimated soil property value. For our project, the main challenges is not how to assign a measure of uncertainty to each predicted value but rather how to establish procedures and mechanisms to help users make use of these measures of uncertainty in their analyses.

Simon Cox JRC (seconded from CSIRO)

Simon Cox gave a presentation via Skype on the need for open standards and on lessons learned from the geoscience and hydrology communities. This presentation included background information on the reasons why the geoscience community decided that they needed to establish standards for data models for specific disciplines. They wanted scientists in the scientific communities to define their own domain data models before such models were defined for them by specific software vendors. They also wanted to ensure that open discussions within the appropriate scientific communities led to open standards that were widely vetted and accepted.

Simon reviewed the options for governance and formal approval of standards that were considered for the geology community (GeoSciML). Simon recommended adopting a governance structure that received its authority from a scientific society affiliated with the International Commission of Scientific Unions (ICSU) as opposed to an organization such as ISO that is essentially concerned only with setting standards and has little or no appreciation of the underlying science. Establishment of a governance structure under a scientific society supports the idea that the setting of standards is important and should be taken seriously.

Simon reviewed the need for open standards and the importance of not reinventing anything that was already invented or produced by someone else. The basic idea of standards for a geoscience community (GeoSciML) is to define a standard for the exchange of similar, but often not identical, information between regional partners in a global scientific community. Initially, the focus is on defining a common understanding of the conceptual models that underlie the science for a particular domain. Using the example of the geology community, Simon listed basic concepts such as map features, geologic features, geologic structures and geologic units. A basic pattern is the capture of information as conceptual entities that can be linked to spatial entities such as points, lines, curves, areas, pixels or volumes. For example a map polygon is a feature that can be linked to a legend. A second basic pattern of science standards is identification and documentation of a controlled vocabulary that can be agreed upon to describe the features of interest. A controlled vocabulary is often captured in a data dictionary. In setting standards, agreement has to be achieved on both the definitions and attributes of the conceptual entities to be described and on the controlled vocabularies to be used to describe them.

Simon provided comments on some of the mechanics that can go into setting standards. He described and pointed out the Wiki site that was set up and used by the geology community to facilitate interactive collaboration in the discussions that went into devising the GeoSciML standards.

www.seegrid.csiro.au/twiki/bin/view/Main/WebHome

The GeoSciML collaboration involved a large amount of work and a large amount of collaborative discussion. Test beds were set up to structure collaboration on design and testing of the conceptual models required to support geological descriptions. Test beds involved a complete set of participants interacting to describe standards that permitted data and maps to be delivered and exchanged between diverse sources and clients. These were not just visual maps but complete XML data sets with reusable data. The collaborating scientists met at least twice per year between 2004 to 2009 to review progress and formulate plans for moving forward. These meetings were not funded and virtually all effort was based on contributions of time and resources provided by the institutions and organizations that the participants worked for (mainly geological survey organizations). Information on all meetings, documentation and test bed materials was stored on the Wiki site for examination and comment by all interested contributors.
Simon suggested that the *GlobalSoilMap.net* project could learn and benefit from the experiences and work of previous activities and would not likely require as many meetings or test beds. Many of the concepts and definitions already recorded for the GeoSciML are "free" for use and reuse by the *GlobalSoilMap.net* project.
Simon explained that the GeoSciML initiative was organized under the Open Geospatial Consortium (OGC) because this was seen as a neutral body whose effectiveness was not hampered by jurisdictional rivalries or conflicts. OGC is a mixture of industry (Google, ESRI, Oracle) and government (NGIA, EPA, INSPIRE, ESA, JRC) that has been around for about 15 years and that was initially funded and promoted by the US Defence Department (NGIA) but that has become increasingly independent.
Simon then went on to give a brief explanation of the difference between UML and XML.
UML is an object oriented modelling language. It is essentially an elaboration of entity relationship (ER) modelling that has been prevalent in the data base field for decades. It provides tools and a vocabulary for describing conceptual objects for a domain of interest.
XML is a mark up language with similarities to HTML. It is used to facilitate exchange or transfer of data. Extended to the geographic domain (GML), it is used to exchange spatial data.

Jim Fortner (USDA: NRCS , NCSS)

Jim provided an introduction to the US experience in collecting and harmonizing soils data with an emphasis on experiences in providing access to digital data to end users over the web. Of note was the level of effort that went into designing and maintaining the cyberinfrastructure services. About 23 people contributed to the cyber-infrastructure component of the National Cooperative Soil Survey (NCSS) with 3 database administrators, 10 programmers and 10 domain or business experts. He discussed the governance and organizational issues required to achieve and maintain standards and consistency in applying the standards among the state participants in the NCSS. The need was discussed for business requirements studies to inform data model and data base design and for published standards to establish and enforce consistency. The US NCSS soil data base is about 250 GB in size with about 1/3 of it being each of

spatial, attribute and interpretive data. The US is intending to migrate their spatial data services from ESRI IMS to SQL Server for reasons of both licensing and efficiency.
A well designed data model and data base needs a well documented set of business requirements. We need to identify how we expect users to interact with, and use, the data. To do this, we need to collect and document information on end users (use cases). We require substantial and dedicated resources to develop and maintain a cyberinfrastructure to provide users with interactive access to the data in a manner than is meaningful and useful to them.

Simon Cox JRC (seconded from CSIRO)

Simon Cox returned via Skype for a presentation in which he elaborated on UML as a powerful tool for capturing and describing concepts used by a scientific community to describe the objects it studies. Simon elaborated on the observations and measurements (O&M) model as a mechanism for developing XML model schemes. An "observation" is "an act that involves a procedure applied at a specific time". A "result" is "an estimate of some property". A rule of entry for developing standards that are compliant with existing standards and regulations is that we must use UML methods. There is a need to develop a soil feature type catalogue and publish it. To do this, we have access to, and can make use of, reusable concepts and code that has already been produced and documented by others. On the Wiki site developed and maintained for the GeoSciML working group there are resources and tools that we can use. We should develop expertise and competence in using O&M modelling and UML within the soils community and within the *GlobalSoilMap.net* project.

Day 2: Wednesday 2nd December (Chair Peter Wilson)

Sonya Ahamed CIESIN

Sonya reviewed the presentations and discussions of the previous day. She noted that two to four conferences per year were needed to coordinate data test bed activities for GeoSciML. She reviewed various acronyms and their meanings; specifically UML – an object-oriented entity-relationship modelling approach; XML – a data exchange standard and GeoSciML a specific type of XML; CIESIN – an ICSU data centre at Columbia University; GfE – Geoinformatics for Geochemistry; and SEDAC – the Centre for Socio-Economic Data Applications. She noted the need to obtain an International Soil Sample Number to meet standards as discussed in Seattle in September 2007.

Bob MacMillan ISRIC

Bob went over a list of elements that could be considered in discussion the scope of the *GlobalSoilMap.net* project in terms of content. This list identified components related to evidence (legacy point and map data, new data and existing knowledge), covariates (e.g. DEM derivatives, satellite imagery), prediction methods, intermediate outputs and final outputs as per the specifications document. The question was raised whether all listed elements ought to be included within the scope of the design and implementation of the data model. In particular, should design and implementation of the data model include the ability to store and process covariate data sets, prediction methods and intermediate data products?
The specifications that have been agreed for the *GlobalSoilMap.net* project were discussed. There is a need to define and publish a base grid to establish the precise locations for which the *GlobalSoilMap.net* project will produce predictions at a 3 arc

second spacing. From a data model perspective, the advantage of specifying a base grid is that delivery of results can be tied to identification codes for specific locations and no spatial data (coordinates) need to be passed to the data base by data producers. A base grid will also ensure that all data delivered by any data producer will fit seamlessly, and without any offset, within a final spatial database.

The discussion of the scope of the data model also led to a review of the work plans and objectives of the AfSIS project to determine which activities, related to the data model, had been expected to be completed by each of the parties in the consortium. The decision was made to increase communication between the concerned parties to better understand and coordinate activities related to development of the cyber-infrastructure required to support the *GlobalSoilMap.net* project.

Each of the attributes that have been proposed to be collected for legacy soil profile data for AfSIS were discussed. A "short list" was defined to be collected for each soil profile as well as a "long list" of a relatively complete set of attributes that could be collected for each site. It was recommended that the *GlobalSoilMap.net* project restrict its activities to entering only data from the "short list" but that local partners be provided with capabilities to enter the full range of "long list" data should they wish to find and allocate the resources to do so. The short and long lists were produced and provided to Sonya as information for producing example data bases and entry screens for entering AfSIS soil profile data.

Peter Wilson CSIRO

Peter led a discussion and modelling exercise in which the UML Observation & Measurement approach was used to identify and describe the basic entities the *GlobalSoilMap.net* project would need to define. There is no need to define a common data model for all data used to produce the final products for the *GlobalSoilMap.net* project. The *GlobalSoilMap.net* project only has responsibility for collating output data and that the various nodes had responsibility for actually collating and analysing the data sets required to produce the predicted soil property values. As such, the *GlobalSoilMap.net* project only has a need to be able to ingest, store and deliver output products of grid maps with soil property values.

Amir Pourabdollah University of Nottingham for eSOTER

Amir gave a presentation of the SoterML data model for storing information about soil polygons and the soil individuals (profiles) described as components of each polygon. The model was presented as complying to all XML standards but Simon Cox identified a number of critical aspects that he felt were at odds with XML standards. It might be only a matter of a few hours to make the SoterML model fully compliant with XLM standards.

Peter Wilson CSIRO

Peter gave a presentation prepared by Rob Atkinson and David Lemon to illustrate the concepts of applying a UML approach of "observations and measurements" to describe simple conceptual entities. The presentation illustrated how different descriptions of similar things (a car) could be abstracted and entered into a common data model that described an agreed-upon subset of the attributes of all cars.

Day 3: Thursday 3rd December (Chair Jim Fortner)

Sonya Ahamed CIESIN

Sonya reviewed the presentations and discussions of the previous day. She clarified the role of the UML approach in reconciling various different data models (e.g. ASRIS, SOTER).
The discussion considered how to use the enterprise architecture (EA) tool and "Rick's tool" to conduct a XML modelling exercise for soils information. The common UML data model from such an exercise can become a candidate for implementation as a universal or common data base. The common UML model at a minimum permits exchange of a subset of the total range of soils information of interest. Three main actions that needed to be undertaken were (i) linkage of the activity to the GEO and GEOSS process, (ii) development and application of UML standards to define the data model through a series of test bed projects and (iii) identification and implementation of a process for requesting and obtaining an international soil sampling number for each soil site and sample.
A discussion followed on the scope of the data model and whether an attempt should be made to capture the full spectrum of inputs and outputs, including the procedures used to produce predicted outputs. Tools for capturing scientific work flows, such as Kepler, from San Diego State and Trident from Microsoft, were discussed in the context of how the entire process of applying predictive methods to input data layers could be recorded and made reproducible. Some tools may not be able to capture all aspects of the scientific work flow. There was concern that the problem of defining a data model big enough and robust enough to permit storage and analysis of the full range of evidence, covariate, model rule and output data would prove to be complex and time consuming. Using the data sets assembled for the "proof of concept" projects could be used as vehicles for developing test bed cases for defining common XML standards for global soil information.
Progress in developing the XML model be decoupled from activities required to mode the project forward in the sort term.

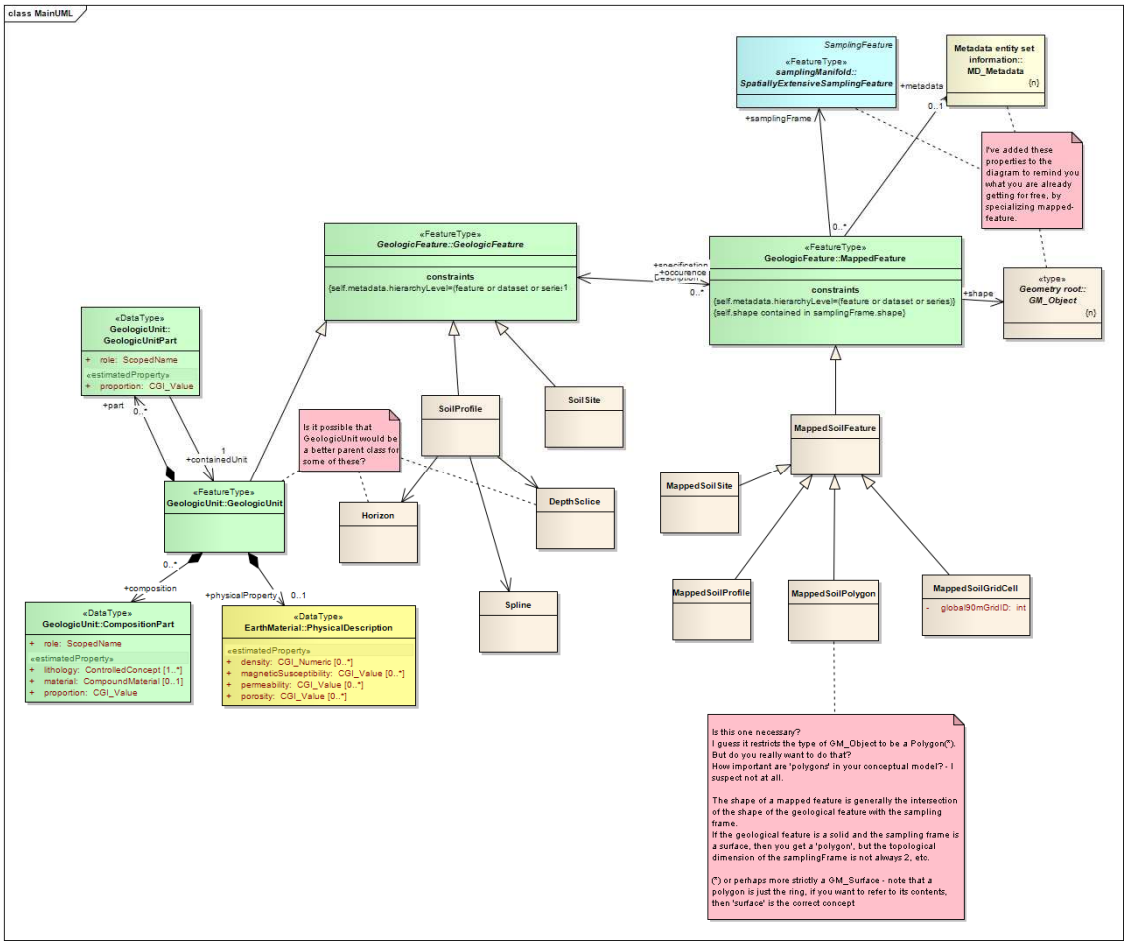
Rob Lokers WUR

Rob gave an outline of his experience in information modelling and described two recent projects. For this project, he noted a conflict between a need to respond to short term operational requirements and a need to develop a complete and robust UML data model in the longer term. A well designed data model was suggested for use in the short term with the expectation that it would fit into a common UML model developed in the future. He suggested that we not mix the test bed activities required to develop a comprehensive universal soilML with short term operational activities. When a UML is developed to permit transfer of legacy data between two or more data models there is still a need to develop a common model and there is often a need to develop an implementation of the common model to hold and store data from diverse sources.

Peter Wilson CSIRO

Peter began to apply the EA concepts to define the main soil feature types that would be required to hold and distribute the information that the *GlobalSoilMap.net* project has committed to produce and distribute (as per the specifications). A preliminary UML

class diagram of feature types and their properties was prepared and sent to Simon Cox for review and comments were received.



(Draft model with Simon's comments)

Day 4: Friday 4th December (Chair David Jacquier)

Sonya Ahamed CIESIN

Sonya reviewed the presentations and discussions of the previous day. The discussions of Thursday had led to a decision to address the AfSIS data capture needs by developing a narrow entry screen that used a hard coded structure to capture each unique source of soil profile data. At the same time, a UML compliant data model will be defined and procedures will be developed and demonstrated to ingest data entered according to the hard coded entry modules. Sonya is to undertake this task with cooperation from ISRIC and Bob.

Capture of legacy map data will be addressed as a separate topic with a data model to be built that can reside on top of the data model defined for soil profile data. There was discussion and recognition of the need to increase communication and coordination between ISRIC and CIESIN, in particular with respect to planning for and implementing the cyber-infrastructure required to support both short and long term project needs. There needs to be a clarification of roles and of budget and task allocation as per the signed contracts. Sonya indicated that the SIESIN budget for supporting cyberinfrastructure had been cut by 50% and that some tasks specific to AfSIS needs had

also been reallocated resources originally targeted at supporting the global infrastructure needs.

This led to a review of general planning and budgeting for *GlobalSoilMap.net*. It was proposed that the task of running test bed exercises make use of data collected to support the proof of concept mapping activities planned for January-may, 2010. It was proposed that the workshop planned for May to review proof of concept results also include a period for discussion of data model test bed activities and progress. It was proposed that the project encourage node leaders to set target dates for being ready to move into operational mapping. Ideally operational mapping could begin as early as September-December, 2010.

The rest of the morning was spent in reviewing the discussions and decisions of the previous 3 days and producing a targeted list of specific follow-up actions and activities. These appear at the beginning of this document.

Willem observed that there was a dichotomy with the group being confident about some aspects and confused about others. He felt that there was confidence about conceptual aspects of the soil data model but confusion about details related to implementation. A group of 10 or more people was too large to deal effectively with details of implementation and that these details should be dealt with by a smaller task group. He felt there was a need to separate short term operational goals from longer term development of conceptual models and XML data models.

Bob identified a number of issues that had not been dealt with in any detail during the course of the workshop and identified these as “loose ends” that needed to be addressed in the coming months. Discussion and selection of a protocol and work flow for first rescuing and then converting or harmonizing legacy map data in Africa and elsewhere was a key topic that was not discussed. Other loose ends were planning and installing the cyber infrastructure needed to support both short term and longer term needs and developing a design and plan for delivering the final products of *GlobalSoilMap.net* that was informed and directed by user needs as elaborated by use case studies.

Prem gave an overview of plans to support global soil information needs within ISRIC and thanked all the participants for attending. The workshop adjourned at approximately 13:00 hrs.