Development of Soil Metadata Standards for International DNA Sequence Databases

James R. Cole[A], David D. Myrold[B], Cindy H. Nakatsu[C], Phillip R. Owens[C], George Kowalchuk[D], Christoph Tebbe[E], and **James M. Tiedje[A]**

[A]Michigan State University, Center for Microbial Ecology, Plant and Soil Science Building, East Lansing, MI, 48824, colej@msu.edu; tiedjej@msu.edu
[B]Oregon State University, Department of Crop and Soil Science, 3017 Agriculture and Life Sciences Building, Corvallis, OR 97331-7306, david.myrold@oregonstate.edu
[C]Purdue University, Department of Agronomy, 915 West State Street, West Lafayette, IN 47907-2054, cnakatsu@purdue.edu; prowens@purdue.edu
[D]Netherlands Institute of Ecology, P.O. Box 40, 6666 ZG Heteren, The Netherlands, G.Kowalchuk@nioo.knaw.nl
[E]vTI - Institut für Biodiversität, Braunschweig – Germany, christoph.tebbe@vti.bund.de

Abstract
The considerable growth in DNA sequencing and its application to microbial communities in many environments, including soils, has drawn researchers from many fields to soil microbial ecology and genomics studies. International convention is for all sequence data to be deposited in public databases as a community resource. However, for this data to be broadly useful information about the data, i.e. metadata, should also be deposited along with the sequence data. A subcommittee of Terragenome, a recently established international consortium to facilitate cooperation on soil metagenome studies, has developed a set of soil features important for understanding soil biology and for interpreting the sequence data. The committee has also defined an associated controlled vocabulary that will allow scientists to search the databases for sequence information that correlates with user defined environmental attributes. The work of the committee is summarized here and has been submitted to the Genomic Standards Consortium (GSC) so that it can be harmonized with data standards for other environments and be implemented as a GSC MIMS standard for use by the global scientific community.

Key Words
Metagenomics, Soil metadata, DNA sequencing, Microorganisms, Ribosomal RNA.

Introduction
Metagenomic studies generate huge amounts of sequence data; however, it is important to realize that these data are useful only within the context of the sample itself. This context is provided by metadata, which includes information necessary to repeat sample collection and preparation, and that describes the key factors that determine the composition of the microbial community and its function. Our efforts are extensions of metadata standards developed for ecology (Michener *et al.* 1997) and more recently for genomics to create minimum information checklists (http://darwin.nox.ac.uk/gc_wiki/index.php/GSC).

Methods
The attendees at the Terragenome II meeting in Lyon, France, in December 2008, endorsed a process of surveying the international community of soil biology researchers for the soil features that they thought would be important to understanding soil microbial community structure and activity, and that would be important to interpreting the molecular data entered into GenBank, EMBL and DDJB. This Web-based survey was available for respondents for approximately 4 months in the spring of 2008. Email address lists from several sources, including the international Terragenome contact list, were used to alert soil biologists of the request. The respondents were asked to evaluate a set of standard soil attributes for both their relative importance to understanding soil microbes and for the degree of difficulty in measuring that feature, the latter important for indicating likely compliance (Figure 1). One hundred five individuals from 15 countries responded. The summary information from this survey is at: http://cme.msu.edu/SoilMetadata/results.html.

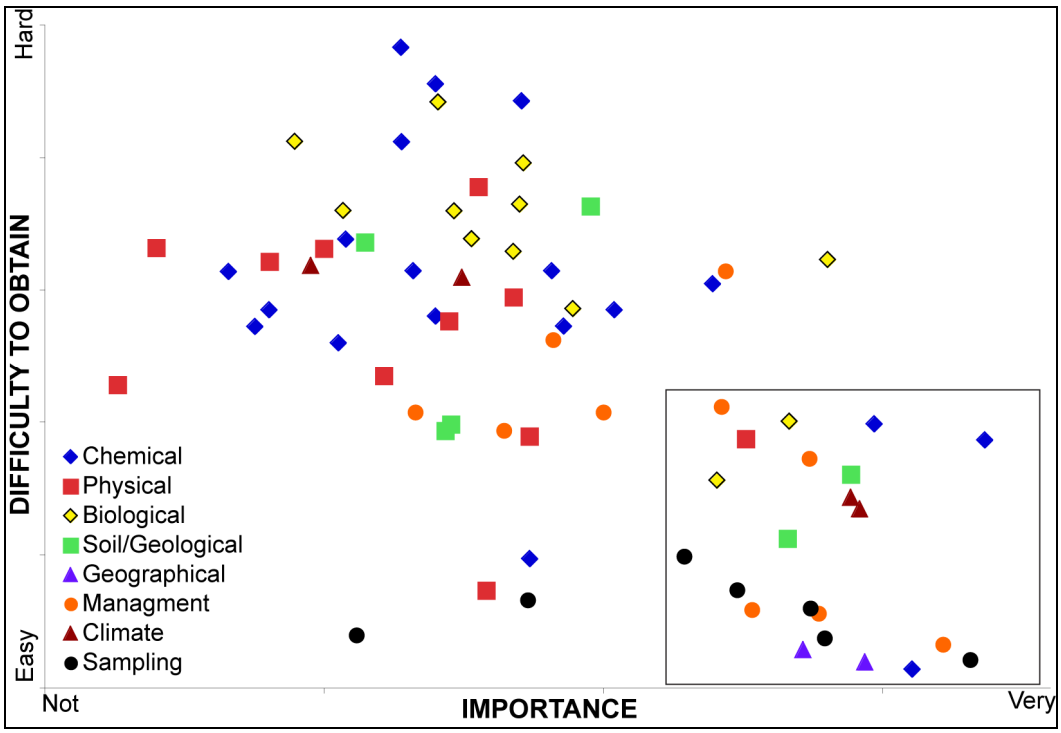


Figure 1. Results of Soil Metadata Survey with Importance and Difficulty displayed. Items in the box were the high priority items of focus in creating the data collection form. This graph can be viewed online [http://cme.msu.edu/jforum/SMSurvey.html] with mouse-over showing a description of each point.

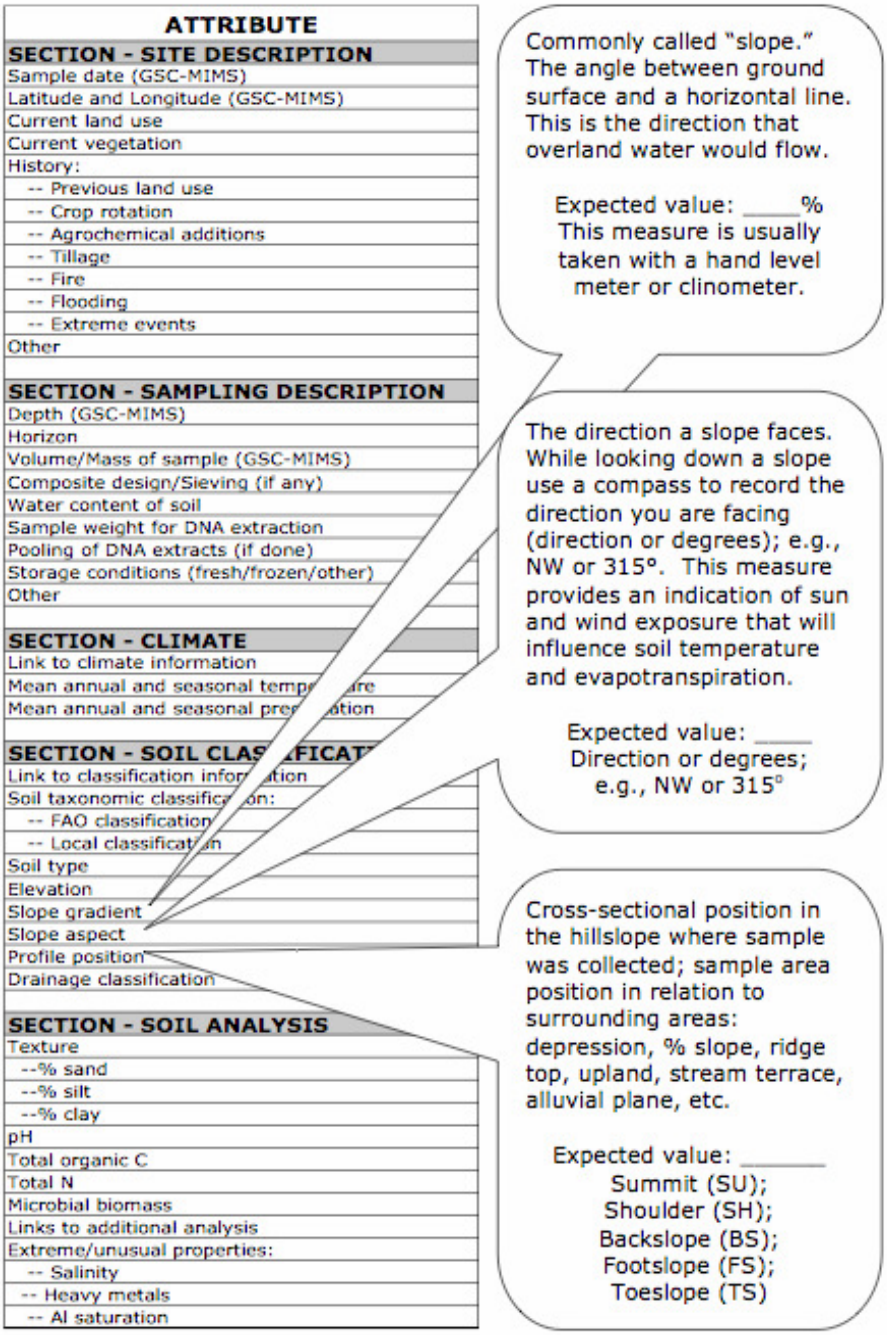
At Terragenome III in Uppsala in June 2009, a subcommittee was formed (the authors on this paper) to take the survey data and refine it into a priority set of features important to soil biology, organize it in a logical manner, develop a controlled vocabulary, provide a set of definitions for those not familiar with soil science terms and submit the recommendations to the international Genomic Standards Consortium. This has been done. The next stage is to work with database entry specialists to implement the metadata entry forms into a convenient and compliant tool. This is underway. The major primary sequence repositories as well as the metagenomic specialty databases such as CAMERA, MG_RAST and IMG/M will also be hosting these metadata.

Results

The structure of the proposed soil metagenome metadata consists of: (1) site description, (2) sampling description, (3) climate, (4) soil classification, and (5) soil analysis. Ideally, the metadata should be important and informative, easy and inexpensive to obtain, and collected by established and standard methods. The elements of the currently recommended data form are illustrated in Figure 2.

The site description documents the location and situation at the time of sample collection and also provides the information necessary to relocate the sampling site. The sampling description provides details related to the collection and processing of the sample, which would be required to repeat the analysis. Also included in these two categories is descriptive information related to the past and current status of the site, which may be useful in interpreting the metagenomic data or ensuring that similar conditions exist if a site is re-sampled.

Figure 2. Soil attribute form. Users can mouse over the attribute name and view a pop-up with the definition and expected value (accepted units). Drop-down selections, such as illustrated, provide the primer for features less known to the diverse user community.



The remaining three categories provide data about the environmental and edaphic factors that are likely to influence the composition of the microbial community. Environmental factors include basic information about the site's climate, which can likely be obtained from the historical records of nearby weather stations. Soil classification and analysis were separated based on the idea that the classification information is integrative in nature, reflecting the soil forming factors acting at a given site. They are descriptive (i.e., qualitative) in nature. The soil analysis data are quantitative and reflect soil properties that are likely to be most significant regulators of microbial community composition.

A primer was also developed to aid this diverse community of scientists we expect will be depositing soil metadata. There were multiple purposes for the primer. (1) To recognize the international differences in terms and definitions used in soil science. (2) To provide some guidance on the expected metadata to be deposited. (3) To provide some simple definitions for those who have not had formal training in soil science. (4) To provide some guidance on measures that can be taken at the time of soil sampling. (5) To provide

some web resources that can be used to obtain more in depth information if desired. The primer was adapted from information published by two major sources, the Schoeneberger *et al.* (2002) soils field guide and U.S. Department of Agriculture-Natural Resources Conservation Service (USDA-NRCS) soil survey manual. Information was cross-checked with any available soil resources from other countries and international societies.

The major topics covered in the primer are soil property measures and land use descriptions. We expect most of the metadata will be easily deposited using drop-down menus, such as illustrated in Figure 2 for a particular item. The soil properties defined in the primer are: slope, drainage, horizon, and texture. Definitions are also provided for land use in general and specific agricultural tillage treatment of soils. The current data form is at: http://cme.msu.edu/SoilMetadata/SoilMetadataEntryTEMPLATE.xls

Conclusions

Metadata will help place annotated soil metagenome data into their correct context. By submitting these standards to the GSC for potential inclusion in the MIMS (Minimal Information about a Metagenome Sequence) standard (Field *et al.* 2008), these standards will reach a larger audience and be incorporated in data preparation and search tools supporting the GSC standards. Researchers from many fields -- e.g., marine biology, human microbiome, environmental engineering, as well as soil science -- will be able to query and retrieve metagenome data from all studied microbial communities based on their metadata content. This will facilitate comparative studies and allow future researchers to incorporate today's data into future integrative studies and advance our understanding of soil microbiology in ways that we may not be able to foresee today.

References:

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, dePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A (2008) Towards richer descriptions of our collection of genomes and metagenomes. *Nature Biotechnology* **26**, 541–547.

Michener WK, Brunt JW, Helly J, Kirchner TB, Stafford SG (1997) Non-geospatial metadata for the ecological sciences. *Ecological Applications* **7**, 330–342.

Schoeneberger PJ, Wysocki DA, Benham EC, Broderson WD (2002) Field book for describing and sampling soils, Version 2.0. Natural Resources Conservation Service, (National Soil Survey Center, Lincoln, NE).

U.S. Department of Agriculture, Natural Resources Conservation Service. National Soil Survey Handbook, title 430-VI. Available at: http://soils.usda.gov/technical/handbook/ accessed [09/22/2009].