12 June 2017

How do we get the most out of soil data? The opportunities and challenges of developing open soil data

Marcia DeLonge[1], and Nathaniel E. Seavy[2]

1. Union of Concerned Scientists, 1825 K. St. NW, Suite 800, Washington DC 20012
2. Point Blue Conservation Science, 3820 Cypress Drive, Suite 11, Petaluma, CA 94954

This paper was produced with financial support from the TomKat Foundation. The paper benefited from conversations with participants at workshops on soil carbon held in Bismarck, North Dakota in July 2016 and Pescadero, California in February, 2017.

-

Climate change has refocused several academic and practitioner communities within the agricultural sector on the importance of soils, for both adaptation and mitigation purposes (Paustian et al. 2016). This soil renaissance has occurred at a time when advances in both democracy and technology are changing the way that we do science. Today, there is a greater emphasis on data sharing in both arenas, alongside a growing recognition that science can be reformed to encourage researchers to work together, rather than focusing on individual achievements (Casadevall and Fang 2012). These philosophical and cultural shifts are associated with technological advances that make it easier and more effective to share large volumes of data across many machines, quickly and efficiently distributing information to a vast audience (Kambatla et al. 2014). In particular, the related technologies have facilitated the move to toward open access data that has been occurring across the natural sciences, from genetics to ecology (Reichman et al. 2011). We propose that our understanding of soils and soil carbon, particularly in light of their associated complexities and uncertainties, can benefit from these developments. In this paper, we review some of the advantages and challenges of making soil data open and accessible, as well as some opportunities and tools for moving forward.

Benefits and challenges of open and accessible data

Science can progress faster when data are open and accessible. Open data allow results to be verified and facilitate additional or expanded analyses as new hypotheses, methods, and information become available (Pampel and Dallmeier-Tiessen 2014). Furthermore, in cases where research is publically funded, it has been argued that the findings fundamentally belong in the public domain (NSF 2016, USDA 2016). Lack of transparent and available data has also been a barrier, for example, in assessing the effectiveness of government conservation programs (Rissman et al. 2017, Ristino and Steier 2016). These points may explain several recent efforts to improve the availability of data created by scientific institutions, including public research organizations, private foundations, and journals (Gewin 2016).

At the same time that open data has tremendous benefits, it may also have real or perceived downsides for scientists, individual landowners, and regulators. For scientists, data sharing can present a risk of others taking credit for or misusing hard-earned observations (Gewin 2016). For this reason, data sharing oftentimes occurs informally among trusted colleagues, and scientists are more willing to share data if credit and first rights to publication are retained (Wallis et al. 2013). When data are collected on private property, landowners may also be at risk of negative consequences if data sharing makes

producers less competitive or more vulnerable to regulation (American Farm Bureau 2016, Stubbs 2016). While many farmers recognize the potential benefits of big data, concerns about ownership and use of agricultural data are growing (Carbonell 2016, American Farm Bureau 2016, Stubbs 2016). Finally, regulators may be concerned about open data if public information on the location of monitoring plots could make those plots vulnerable to tampering. Data sharing with consent as a prerequisite and data anonymization are broadly applicable strategies that have been successfully adopted to address tensions between public benefit and individual privacy in the health care domain (El Emam et al. 2015, Tucker et al. 2016). These examples may offer helpful insights that could be used to design systems, policies, or protocols that harness the potential advantages of open data within the natural sciences, while preventing many of the potential consequences.

Data sharing in the soil sciences

From a soil science perspective, both consent and anonymization have been practiced. On one end of the spectrum, the Soil Carbon Challenge (http://soilcarboncoalition.org/challenge) requires that all participants consent to sharing their data openly. On the other end, the Natural Resources Conservation Service (NRCS) has strict guidelines in place to assure anonymity (NRCS 2009). Anonymization of soil data is usually accomplished by eliminating any identifiers that tie a data record to an individual (e.g., names, phone numbers, or email addresses) and reducing the accuracy of the geographic location (e.g., reporting only the county where a sample was collected rather than the actual latitude and longitude).

Regardless of which approach is used, having a clear protocol and consensus on data sharing policies is essential (EDF 2016). To advance soil science most quickly, an ideal level of data accessibility would be full open access, where data are available to anyone who is interested and can be distributed with geographic locations. However, to give landowners and other parties the level of protection they may need or desire, other options may be sufficient. For example, data that are open access anonymous could allow landowners to grant permission for a third party to collect data, generate a data set with additional characteristics (e.g., slope, soil series, etc.), and then share the data after the specific geographic coordinates and individual identifiers have been removed. A third option that would still provide an opportunity for accessibility would be to collect data that is shareable on a case-by-case basis with consent and restrictions. Permissions to use data may or may not include geographic locations or other individual identifiers.

From a science perspective, the more anonymous the data, the more challenging analysis may be. For one thing, a landowner could agree to share data on the condition that the final publication did not include the specific location of the data, and may not consent to making those data available to other scientists, potentially hindering the line of research. Another challenge is that soil data, and some of the methods associated with their analysis, are inherently spatially complex (Hengl et al. 2004). As a result, it may be difficult to conduct meaningful analyses without accurate spatial locations associated with each sample, and masking the location can create severe limitations on research. In such cases, sufficient meta-data that provides the context for the measurements is particularly important (Gerstner et al. 2017). For soil information, several pieces of information on slope, aspect, and elevation will be particularly important especially if the spatial location is not available. Additional information on sampling methodologies and land management may also be necessary to provide sufficient context to make the data relevant.

Data aggregation: Sharing protocols, crowdsourcing, and building networks

In addition to making data open and accessible, there can be tremendous benefits to proactive efforts to organize and aggregate data, particularly for soils and soil carbon (Arrouays et al. 2014). These efforts may entail anything from sharing protocols to crowdsourcing to building networks that connect people. The aggregation of people, protocols, research projects, and data into coordinated networks is being applied across the scientific domain, from citizen science programs to nuclear physics (Adams 2012).

One of the exciting aspects of new data-sharing networks has been to connect professionals and amateurs in a manner that engages a diverse audience and accelerates science. For example, promoting engagement earlier in the process – before data is even collected rather than only after data is processed – can improve the consistency, quality, and utility of data. In particular, sharing data collection and methodological protocols can take individual efforts and strategically synchronize them in a manner that allows much larger questions to be addressed. In addition, crowdsourcing efforts can work to mobilize individuals to fill data gaps and exchange knowledge, either with or without shared protocols (Paustian 2013). Crowdsourcing, and the related concept of "citizen science", has recently received new attention from the U.S. Federal Government, indicating its growing importance in science and policy (www.citizenscience.gov; OSTP 2015). Ultimately, aggregating ideas and data may be easiest and most effective through networks. Fortunately, in soil science, networks are beginning to gain traction among both scientists and practitioners (Table 1).

Table 1. A list of some soil data networks and a brief description of their data sharing policies. Our examples focus on networks that specifically address soil carbon, but many networks for other soil characteristics exist and provide examples of collaboration and data sharing.

Network	Data sharing policy
Rangeland Monitoring Network (www.pointblue.org/rmn)	Data collected on private lands can be shared upon request if they are anonymized or can be made available with locations with consent of the landowner.
Soil Carbon Challenge (http://soilcarboncoalition.org/challenge)	All data are open and available on-line.
International Soil Carbon Network (http://iscn.fluxdata.org/)	Data are open and available on-line except that locations of sensitive sites are generalized to 0.1 degrees of latitude and longitude.
Soil Carbon Network for Sustainable Agriculture in Africa (http://reseau-carbone-sol-afrique.org/en)	This network focuses on connecting researchers and generating products, rather than archiving and sharing data.
Permafrost Carbon Network (http://www.permafrostcarbon.org/)	This network focuses on connecting researchers and generating products, rather than archiving and sharing data. Products include spatial layers of soil carbon that are available on-line (http://bolin.su.se/data/ncscd/).

Conclusions

As soil science moves to the forefront of conversations about ways to mitigate and prepare for human-induced climate change, there is a renewed sense of urgency around gathering the data that is needed to identify the best opportunities for management to improve soil health. To the degree that these data can be made openly available and shared, learning will be accelerated. However, the concerns of private landowners must be considered in this process. Already, soil networks are addressing these concerns to responsibly make soil data available to a growing audience of users.

Citations

Adams, J. 2012. Collaborations: The rise of research networks. Nature 490:335-336.

American Farm Bureau. 2016. http://www.fb.org/tmp/uploads/BigDataSurveyHighlights.pdf

Arrouays, D., B.P. Marchant, N.P.A. Saby, J. Meersmans, C. Jolivet, T.G. Orton, M.P. Martin, P.H. Bellamy, R.M. Lark, B.P. Louis, D. Allard, and M. Kibblewhite. 2014. On soil carbon networks. Pages 59-68 in A.E. Hartemink and K. McSweeney (eds.), Soil Carbon. Progress in Soil Science. Springer International Publishing Switzerland.

Carbonell, I.M. 2016. The ethics of big data in big agriculture. Internet Policy Review 5(1). http://policyreview.info/articles/analysis/ethics-big-data-big-agriculture

Casadevall, A. and F.C. Fang. 2012. Reforming science: methodological and cultural reforms. Infection and Immunity 80:891-896.

Environmental Defense Fund [EDF]. 2016. Farmer Network Design Manual. https://www.edf.org/sites/default/files/farmer-network-design-manual.pdf

El Emam, K., S. Rodgers, and B. Malin. 2015. Anonymising and sharing individual patient data. British Medical Journal 350:h1139.

Gerstner K, D. Moreno-Mateos, J. Gurevitch, M. Beckmann, S. Kambach, H.P. Jones, R. Seppelt. 2017. Will your paper be used in a meta-analysis? Make the reach of your research broader and longer lasting. Methods in Ecology and Evolution, Early View.

Gewin, V. 2016. Data sharing: An open mind on open data. Nature 529:117-119.

Hengl, T., G.B. Heuvelink, and A. Stein. 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120:75-93.

Kambatla, K., G. Kollias, V. Kumar, and A. Grama. 2014. Trends in big data analytics. Journal of Parallel and Distributed Computing 74:2561-2573.

National Science Foundation [NSF]. 2016. National Science Foundation Open Government Plan Version 4.0. https://www.nsf.gov/pubs/2016/nsf16131/nsf16131.pdf

Natural Resources Conservation Service [NRCS]. 2009. Natural Resources Conservation Service United States Department of Agriculture Acknowledgement of Section 1619 Compliance. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs141p2_002666.pdf

Office of Science and Technology Policy [OSTP]. 2015. Addressing Societal and Scientific Challenges through Citizen Science and Crowdsourcing (Memorandum) https://www.whitehouse.gov/sites/default/files/microsites/ostp/holdren_citizen_science_memo_092915_0.pdf

Pampel, H. and S. Dallmeier-Tiessen. 2014. Open research data: From vision to practice. Pages 213-224 in S. Bartline and S. Friesike (eds.), Opening Science: The evolving guide on how the Internet is changing research, collaboration and scholarly publishing. Springer International Publishing, Switzerland.

Paustian, K. 2013. Bridging the data gap: engaging developing country farmers in greenhouse gas accounting. Environmental Research Letters 8:021001.

Paustian, K., J. Lehmann, S. Ogle, D. Reay, G.P. Robertson, and P. Smith. 2016. Climate-smart soils. Nature 532:49-57.

Reichman, O.J., M.B. Jones, and M.P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703-705.

Rissman, A.R., J. Owley, A.W. L'Roe, A.W. Morris, C.B. Wardropper. Public access to spatial data on private-land conservation**.** *Ecology and Society*, 2017; 22 (2) DOI: 10.5751/ES-09330-220224

Ristino, L., and G. Steier. 2016. Losing ground: A clarion call for Farm Bill reform to ensure a food secure future. Columbia Journal of Environmental Law 42:59-116.

Stubbs, M. 2016. Big data in US agriculture. Congressional Research Service Report 7-5700, R44331. https://www.fas.org/sgp/crs/misc/R44331.pdf

Tucker, K., J. Branson, M. Dilleen, S. Hollis, P. Loughlin, M.J. Nixon, and Z. Williams. 2016. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Medical Research Methodology 16:77.

United States Department of Agriculture [USDA]. 2016. United States Department of Agriculture Open Government Plan Version 4.0. http://www.usda.gov/documents/usda-open-gov-plan-v4.0.pdf

Wallis, J.C., E. Rolando, C.L. Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS ONE 8: e67332.