An Interdisciplinary Approach to Finding and Using Data for Complex Environmental Modelling Problems: A Soil System Example

Graham Dean[1], Victoria Janes Bassett[2], Ross Towe[1], Vatsala Nundloll[1], Jess Davies[2], Gordon Blair[1]

1. Motivation

Environmental processes are often complex, as they involve non-linear interactions between biological, chemical and physical processes across a range of temporal and spatial scales. Soil systems are one example of this complexity. Soils emerge from the interaction between geology, biology (plant and microbial processes), biogeochemistry, hydrology, and climate. Further complexity arises from the feedbacks between soil and human systems. Soils underpin society: they provide the vast majority of our food, the regulate water flows and quality, as a carbon store they are important for climate regulation. Through agriculture, land use change and pollution, human actions over the last have modified soils for centuries. Yet soil sustainability is often overlooked in policy and private sectors [1]. To understand these complex systems, simulation models are needed to integrate the processes that shape soils and influence their sustainability and that of our ecosystems and societies.

To date, much of the modelling of soils has existed within one domain (e.g. either carbon cycle models, soil water flow or quality, soil erosion or formation). We aim to develop a more comprehensive model of the soil system by coupling soil biogeochemistry (N14CP, see [2]) with erosion and hydrological processes, to create a model that can be used to simulate soil functioning and response under changing environmental conditions. The model can then be used as part of an integrated risk assessment process (for an example of an integrated risk assessment approach used in the flood management domain see [3]) to inform decision making, enabling sustainable soil management and ensuring continued provision of vital soil services such as food provision and carbon storage.

Large, or even vast, amounts of data from diverse sources are required to not only act as model input, but also to determine initial conditions, parameter values and to calibrate and validate models. Searching for this data can be incredibly time consuming and takes time away from furthering our understanding within environmental science. Standard methods of data searching can often be unsuccessful and miss sources of useful data. We are interested in the potential of semantic web technologies to enable more efficient discovery and querying of data sources.

This paper addresses the important and under researched area of how to manage data associated with such large-scale integrated modelling experiments. In particular, we are adopting an interdisciplinary approach to determine how digital technologies might contribute to environmental science. The overall goal of the paper is to evaluate the potential effectiveness of semantic web technology in addressing the needs of complex environmental modelling.

[1] School of Computing and Communications, Lancaster University, UK
[2] The Pentland Centre for Sustainability in Business, Management School, Lancaster University, UK

2. Interdisciplinary Approach

We are taking an interdisciplinary approach to addressing the data challenges associated with modelling soil systems, which are shared with many other environmental modelling applications, using collaborative expertise across computer science and environmental science. The technical approach is based upon using innovative semantic web technologies [4] and cloud-based computational resources to address data integration, querying and reasoning across a wide range of different data sources and predictive models.

2.1 Overview of Semantic Web Technologies

The semantic web aims to provide support for an interconnected web of data to allow computers to do more useful work [5]. It is supported by five technology groups addressing various aspects of this data integration:

Linked Data: The Resource Description Framework (RDF) provides a standard model for linking data through meta-data and accessing data over a network.

Vocabularies / Ontologies: Vocabularies allow data to be expressed conceptually from different perspectives and support the integration of datasets from different domains. Technologies such as OWL and SKOS support this data and knowledge organisation.

Querying: Once data is enriched through suitable meta-data it can then usefully be queried (along with other datasets) using a semantic query language. SPARQL (and associated extensions such as GeoSPARQL) support this querying of semantically-enriched data.

Inferencing & Reasoning: Often we want to reason across different datasets with different properties. OWL supports this through the development of class hierarchies and relationship models.

Vertical Integration: Finally, semantic web technologies do not exist on their own. They need to be integrated into domain specific conceptual frameworks and supporting tools. Communities are tailoring existing technologies and building new technologies specifically for their domain of interest.

2.2 Data Needs

For new environmental modelling projects, a significant amount of time is typically spent finding data to satisfy specific requirements. This includes model calibration and validation data, and data required to set boundary and initial conditions. Ideally, there should be software support to assist in the identification and retrieval of relevant datasets from a range of sources. In relation to the soil systems model in this project, this approach could be used to search for data to test (e.g. soil surveys and long-term experiment sites), parameterise (e.g. plant type properties) or run (e.g. climate or historical data) the models. Reasoning across heterogeneous datasets and models provides a means of integrating a variety of data sources, such as those with varying spatial and temporal resolutions, which can then be used for model simulation purposes. Similarly, at a later stage in the project these methodologies can be used to analyse large

datasets acquired from model outputs. Finding new high-level abstractions to interrogate this data will provide a more sophisticated means of analysing this data.

Discovery of data is significant, but will not satisfy the range of data-related issues found in environmental modelling. Further challenges exist in bringing together and querying across disparate datasets as well as understanding their provenance and associated uncertainties.

2.3 Connecting Data Needs to Semantic Web Technology

These data needs stated above in Section 2.2. lead us to four high-level themes which underpin the semantic web:

- Semantic Integration of data and models: Different datasets and models often come from different application domains that have diverse understandings and terminology relating to data types. Semantic integration allows us to make these relationships explicit and suitable for machine processing.
- Data and Model Discovery: It can often be extremely time consuming to both find suitable datasets for specific geographical areas and to use a predictive model within available computational resources. We would like to demonstrate the value of making these resources discoverable and usable directly on the web.
- Reasoning: We want to be able to reason across diverse heterogeneous datasets with differing spatial and temporal characteristics.
- Uncertainty Management: Uncertainty is an inherent problem in modelling and different types of uncertainty need to be addressed in an explicit manner and accounted for using appropriate mechanisms.

We are interested in how these four technology themes can potentially help to address some of the challenges found in discovering and utilising data for environmental modelling.

3. Issues in Data Discovery and Utilisation in Soil System Modelling

To illustrate the demands and issues surrounding discovery of data for environmental modelling we identify nine classes of problems. These problem classes were identified through a process of reflection by the environmental modellers who have experience in a wide-range of modelling projects. Some of these issues may relate more to particular types of environmental modelling (e.g. scarcity of data in soil system science), whilst others seem more general in nature.

- *Ease of measurement*: the variable(s) may require manual measurement if it cannot be measured using an automated technology or be observed remotely. This is more labour intensive and therefore more time consuming and costly which could result in a barrier to data collection. A similar issue may arise if high temporal or spatial resolution/frequency of observation is required, but not easily collected.

    - Example: Soil chemistry data – variables such as soil carbon or nutrient contents cannot be measured using an automated process. The observation requires manual collection of soil cores and laboratory analysis.
- *Temporal scale of interest*: some environmental processes are slow, occurring over decadal to centennial timescales. Data sets covering long time periods are valuable in these instances, however, they are often more difficult to find and utilise due to lack of consistent support for long term data collection (e.g. funding and staffing) , particularly where governmental support is not present. Issues regarding archiving of such data can also create a barrier.
    - Example: Soil organic matter turnover occurs over timescales raning from minutes to millennia. Repeated sampling of soil organic matter over long timescales is valuable in calibrating and validating models. Such data exists from long term monitoring sites but these are limited in number and spatial coverage.
- *Diversity in methods of observation*: where multiple methods exist to measure a particular variable, this could create consistency issues within datasets.
    - Example: There are numerous laboratory methods for testing soil phosphorus, not all of which are directly comparable to each other, or to modelled states.
- *Level of regulation*: where there is no regulatory mandate to collect or monitor environmental data (at national or international level) there is often no national facility to collate the data. Such regulation can also be useful by setting standardised practices for methodologies and units of measurement.
    - Counter example: CEH National River Flow Archive – this web based data portal provides coverage of multiple flow gauging stations across England and Wales in a standardised format. This is funded by the UK's NERC (Natural Environment Research Council).
- *Frequency of data use*: where data is infrequently used there is little demand (and incentive) to make the data easily available. This creates a positive feedback as the fact the data is not easily available it is used less.
    - Counter example: CEH National River Flow Archive – this data is easily accessible and used frequently. Therefore, the data facility has secure funding and is maintained.
- *Extreme value errors*: some variables may have issues with the level of accuracy associated with extreme values. This could be linked to a high level of uncertainty in these values, possibly due to fewer observations in extremities.
    - Example: River sediment data during high flows – fewer data exists during high flows, and values are often extrapolated resulting in lower confidence. Hysteresis effects also exist resulting in decreased confidence within sediment volumes.
- *Cultural data sharing practice barriers*: within science and academia there is often no incentive to share data. Culturally researchers may be possessive over data for publication reasons, whilst some data may be of a sensitive nature restricting open access.
- *Data from multiple regions (language barriers):* if data are required from multiple geographical regions it may be held by different organisations making collation more complex. Issues may arise due to permissions for data access, and also language barriers may restrict usage.

    - Example: Long term soil data from experimental sites – for the soil systems modelling we have been collating data from multiple long-term sites across Northern Europe. Collating data from some sites is more complex due to access issues.
- *Trust & Provenance*: subjectivity in observations (e.g. biodiversity) where varying experience may lead to different observational recordings.
    - Example: mixing citizen science collected data with professional survey data.

We analysed each of the issues in the context of a wide-range of scientific and policy stakeholders who have interests in soil system science. Through this analysis, it became apparent that these issues often had an effect on individuals and the scientific community, but the best solutions to these issues would come from the scientific and policy interface, even if the needs of the stakeholders would be quite different.

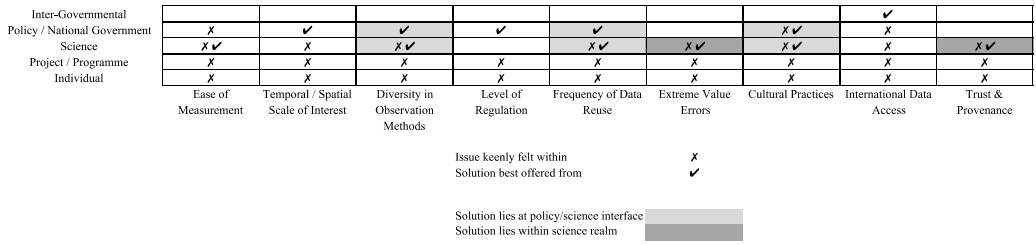

Figure 1: *Analysis of Data Needs*

The data needs described above cover discovery, integration, reasoning and uncertainty management, however, immediate benefits would be felt by addressing data discovery and, conversely, identifying where additional data were required to meet environmental modelling needs. In Section 4 we map semantic web technologies to these two important themes: (i) the need to find data, and (ii) the need to find gaps in the available data.

4. Mapping of Semantic Web Technologies to Problem Areas

In this section, we link the data needs described in Section 3 to the five Semantic Web technology groupings described in Section 2.1. Although there are problems in dataset integration, querying and reasoning and the management of uncertainty, as an illustration we focus on exploring semantic web technologies can help to address data discovery issues.

These mappings below are meant to be simply illustrative to show how semantic web technologies may help address the data discovery data gap problems found within environmental modelling; future work will expand on these and address the other problem areas identified.

	Finding Data	Identifying Data Gaps
Linked Data	Supports and promotes reuse of scarce data through community efforts. Can provide links to alternative language translations.	Identify metrics of coverage from graph analysis. Idealised data models showing empty areas where data does not exist.
Vocabularies / Ontologies	Need to reflect the multi-faceted nature of soil systems. Spatial and temporal attributes required. Make visible observational methods used in data collection. Express assumptions and incompatibilities across different environmental domains.	Ontologies typically expressed at the scientifc level, but more work needed to clarify the needs of, e.g., scientific programme managers and policy officers. Ontologies need to accomodate a wide range of stakeholders with different perspectives.
Querying	Support for finding datasets with appropriate temporal and spatial scales. Identifying appropriate datasets for reuse (e.g. find data from experiments using soil types that are 'similar' to those found in catchment X, not just GeoSPARQL like spatial queries).	There may be gaps in data from a temporal, spatial or uncertainty perspective. An open question is how does one query for gaps in data? Is it implicitly (return of incomplete data sets) or explicitly (query for missing data)
Inferencing / Reasoning	Bringing together process-based and data-based models and understand their assumptions and any compatibility issues.	Does sufficient data exist to quantify the uncertainty around model predictions. Can we adequately reason about the different potential sources of uncertainty?
Vertical Integration	Integration into domain specific toolsets and conceptual frameworks, e.g. GIS tools, risk analysis frameworks, agricultural sustainability models	Tool support required for investment decisions covering scientific programme and policy areas. Needs to provide evidence to support a broad range of data generating interventions.

Figure 2: *Semantic Web Support*

It is notable that much of the semantic web research is focused directly on data discovery issues, but many data issues that the soil science community face actually relate to identifying missing data. Addressing this need will open up interesting research avenues for the semantic web community and further promote the need for interdisciplinary research.

5. Discussion and Conclusions

We have identified a range of data discovery and utilisation problems found in a current soil systems research project. Whilst some of these problems may be specific to soil science, others are more generally applicable into other areas of environmental modelling. These discovery and utilisation problems affect a wide-range of stakeholders, from individual researchers to inter-governmental organisations.

Semantic web technology can help with both discovery of data and linkages across datasets. However, much of this work is focused on science and technology issues and not on policy and governmental issues. Semantic web technology can also help in making data gaps visible to stakeholders and provides evidence for a wide range of interventions from generating simulated data, through directing field campaigns, to scientific programme initiations and policy changes. Although the problems are most keenly felt by individuals, solutions may need to come from other stakeholders.

We have provided a mapping from semantic web technologies to the problems identified i.e. finding existing relevant data and help to expose missing data.

Although semantic web research is active in addressing some of issues we have identified, the area of identifying missing data from temporal, spatial and uncertainty perspectives are less well covered.

Future work will address these issues such as gaps in datasets; making use of our technologies and architecture to automatically simulate missing data on demand as required. In addition to scientific process integration and representation, this project will provide engineering solutions to enable technical integration of model processes and simulation. This will include solutions such as integrating models operating at different spatial and temporal resolutions, potentially coded in different languages and operating in different computational environments.

Acknowledgements

We would like to thank EPSRC grants EP/N030532/1 (Soil-Value: Valuing and enhancing soil infrastructure to improve societal sustainability and resilience) and EP/P002285/1 (The Role of Digital Technology in Understanding, Mitigating and Adapting to Environmental Change).

References

[1] Davies, J.A.C. (2017) The Business case for soil, Nature, 543**,** 309–311, doi:10.1038/543309a

[2] Davies, J. A. C., Tipping E., Rowe E. C., Boyle J. F., Graf Pannatier, E. and Martinsen V. (2016), Long-term P weathering and recent N deposition control contemporary plant-soil C, N, and P, Global Biogeochem. Cycles, 30, 231–249

[3] Harvey, H.; Hall, J. & Peppé, R. (*2012*) Computational decision analysis for flood risk management in an uncertain future, Journal of Hydroinformatics, IWA Publishing, 14, 537-561

[4] Berners-Lee, T., Hendler, J., Lassila, O. (*2001*) The Semantic Web. Scientific American, 284(5):34-43.

[5] W3C (2017): https://www.w3.org/standards/semanticweb/ (07.08.2017)