Many complex resource systems, both natural and human-made, that produce and distribute food, energy and water (FEW), do not exist in isolation. Each resource system may negatively impact how another system is managed, and it is often necessary to find suitable balance and accept trade-offs. For example, water management in California gives first water rights to farmers rather than for human consumption, creating a tension between ecological sustainability and food security. To make matters more complicated, resources are embedded in the complex socio-technical systems where human decision-making is shaped by the goals and capacities of various stakeholders and by the technological capabilities they have at hand (Smith and Stirling, 2010).

Data science is essential in understanding the complex FEW resource systems especially as they come under increasing pressure from changes to the climate. The growing availability of cheap and resilient sensors, proliferation of everyday technologies such as mobile phones, and prevalence of cloud services, creates new opportunities for data collection and analysis. The number and size of the relevant datasets is becoming vast and heterogeneous as well, and with many opportunities come many challenges and dilemmas (Ekbia et al, 2015). To employ innovative data collection and analysis techniques in solving FEW nexus problems, researchers need to make sure that the data and methods they use can be trusted and that the technologies that are being maintained and developed can be sustainably used and deployed beyond the developed world. In what follows we discuss two contributions that data scientists can bring to the FEW nexus problematics, drawing on FEW-related cyberinfrastructure projects at the Data to Insight Center at Indiana University (IU).

1. FEW nexus problems are underexplored in developing regions. The power of data science needs to be extended to those regions to help level the global FEW supply and alleviate the impact of climate change. To be effective, FEW data science techniques, such as modeling and visualization at scale, have to integrate existing technologies available in academia and elsewhere.

The modeling of resource systems relies on two pillars: hardware and devices, and computational analysis. Hardware and devices – sensors, computers, servers, phones, and connecting cables – help to sense, aggregate, and process data and deliver results fast. Computational analysis integrates data from multiple sources, corrects for errors, derives insights from datasets that cannot be processed manually, and presents vast amounts of data in an aesthetically appealing and intuitive way. Many of the skills and resources needed for devices, hardware and computational analysis to run exist in the developed world. The project “Impacts of Agricultural Decision Making and Adaptive Management on Food Security”, awarded by the National Science Foundation (award # 1360463) to Princeton University and Indiana University, extends those capacities to problems of food security in Sub-Saharan Africa, a region that is most vulnerable to food, water, and energy scarcity. The team recognizes the disparities that exist in access to technologies and attempts to remediate them by utilizing three independent technological developments: the rapidly increasing accessibility of cellphones in Sub-Saharan Africa, the decreasing cost of environmental sensing systems, and the creation of cloud computing frameworks that facilitate data transmission and storage.

Current effort is split between the intensive data collection effort in Kenya and Zambia using the three technologies described above and the ongoing development of methods and software that can readily integrate co-collected data on crop growth, rain, planting, and farmer decision-making and generate aggregated results that can be fed back to farmers and policy-makers. The integration methods will incorporate measurements from all three resources, food, water, and energy, to understand how climate variability alters water demands, farmers’ irrigation needs, and subsequent harvesting outcomes in Kenya and Zambia.

Real time data collection for a Sub-Saharan growing season is a unique opportunity to develop more detailed models of community labor sharing, and planting decisions through agent based models. But ingesting real time data into an

1 See TextIt SMS applications platform (https://textit.in/) and its open source option RapidPro (http://blog.textit.in/textit-open-sources-technology-platform-as-rapidpro)
agent-based model significantly increases the model’s size and complexity, so they become too large to run through standard desktop modeling software such as NetLogo\(^2\). The team is working on a customized solution for parallel simulations on high-performance computing resources at IU.

2. Data sharing and re-use are crucial for solving FEW nexus problems. To adopt the culture of sharing, researchers need to trust the data they obtain from external sources. Technology can help to increase the trustworthiness of data by documenting information about data and its lineage.

Data integration and distributed processing require interdisciplinary teams of scientists and technologists who work across disciplinary boundaries, using data and approaches from varied sources and domains. When data and methods are being re-used across domains, trustworthiness is key. The perspective on trust offered below is drawn from a 5-year project “Sustainable Environments Actionable Data (SEAD)” funded by the National Science Foundation (award # 0940824) \(^3\). SEAD responds to the needs of sustainability science researchers, many of whom deal with the challenges within the FEW nexus and need to manage their heterogeneous data (Plale et al, 2013). SEAD provides software to curate data for publishing, reduce the manual metadata markup burden, and publish to a repository of choice.

To strengthen trustworthiness of a complex research object, SEAD employs a simple formalism we call “trust threads” that capture object’s metadata and data provenance (i.e., lineage, see Plale et al, 2011) at the stage when a researcher packages and distributes research data more broadly. We call the time between publication of a data object to when a different researcher accesses the data for reuse the “publish-reuse lifecycle” window. During this window much of what was known about the data can be still captured, and if not captured, it may be lost forever.

The formalism is the basis upon which software is written that implements the model in a controlled and predictable manner. It has the following parts: 1) metadata that describes who contributed to the data collection and processing and what methods were used to process data, 2) states, which define the condition of a dataset as it passes through the publish-reuse window, and 3) relationships that capture how one dataset may be related to other datasets that belong to the same or other researchers. The metadata can be described using any available metadata standard, such as **Dublin Core** or **Ecological Metadata Language** (EML). The states and relationships are drawn from two enumerated sets as follows where relationships are a subset of the properties defined in **PROV-O** ontology language:

- States = \{Live Object (LO), Curation Object (CO), Publishable Object (PO)\}
- Core Relationships = \{wasDerivedFrom, wasRevisionOf, alternateOf \}

Once implemented and embedded in systems that support data sharing, these minimal “trust threads” will allow for the tracking of a provenance “suitcase” enabling the data’s accuracy, validity, and stability to be more easily determined. This, in turn, will enable FEW researchers to acquire and re-use data from more sources and regions, thereby overcoming the lack of reliable measurement-based data and further improving the FEW nexus analytic capacities across the globe (US Department of Energy, 2014).

References


\(^2\) [https://ccl.northwestern.edu/netlogo/](https://ccl.northwestern.edu/netlogo/)

\(^3\) [http://sead-data.net/](http://sead-data.net/)