The ENIGMA Data Clearinghouse: A platform for rigorous self-validated data modeling and integrative, reproducible data analysis

John-Marc Chandonia*,1 (JMChandonia@lbl.gov), Pavel S. Novichov*,1, Adam P. Arkin, and **Paul D. Adams**[1,2]

[1]**Lawrence Berkeley National Lab, Berkeley**; [2]University of California at Berkeley; *co-first authors

http://enigma.lbl.gov

Project Goals: ENIGMA (Ecosystems and Networks Integrated with Genes and Molecular Assemblies) uses a systems biology approach to understand the interaction between microbial communities and the ecosystems that they inhabit. To link genetic, ecological, and environmental factors to the structure and function of microbial communities, ENIGMA integrates and develops laboratory, field, and computational methods.

One of the Grand Challenges of data science is to facilitate knowledge discovery by enabling datasets to be readily analyzable both by humans and by machine learning algorithms. In 2016, a diverse group of stakeholders formalized a concise and measurable set of principles, called FAIR, to increase the utility of datasets for the purpose of knowledge discovery. The four principles of FAIR are Findability, Accessibility, Interoperability, and Reusability. *Findability* means that data are assigned stable identifiers, and properly indexed. *Accessibility* means the data are easily retrievable by people authorized to have access. *Interoperability* means the data are clearly documented using a formal language, in order to facilitate integrated analyses that span multiple datasets. *Reusability* means the data are documented sufficiently well that it may be used by other people than those who originally generated it, and that the provenance of all data is clear.

The latter two principles are particularly challenging, yet critical to achieve, for organizations such as ENIGMA that draw conclusions based on highly integrative analyses of many types of data generated by multiple labs. *Reusability* can be difficult because non-specialized data formats often do not allow or require specification of key details, even basic ones such as units of measurement. As a result, it can be challenging to reproduce or reuse data, because of undocumented assumptions and conventions. Ensuring *Interoperability* between datasets is hard for many of the same reasons: when different teams within an organization produce data, impedance matching must be done in order to perform an integrative analysis. Some sources of impedance are differing units, incompatible scaling or normalization of different datasets, and different identifiers used by different teams to refer to the same objects.

We surveyed hundreds of data types throughout ENIGMA, and discovered that the vast majority of data (from raw assays to processed results) can be represented by a limited number of mathematical data models, such as multi-dimensional arrays of scalars. We believe that this result is generalizable across many fields of research, and indeed, storage formats such as HDF5 and NetCDF-4, along with libraries such as Xarray, are well supported and widely used technologies. However, a common file format alone is not sufficient to ensure adherence to the FAIR principles of *Interoperability* and *Reusability*: in addition, all contents, dimensions, and units in these multidimensional arrays must be formally and rigorously documented.

We developed the ENIGMA Data Clearinghouse, the first general-purpose platform that solves this problem. This relies on three key technologies: 1) to rigorously document context for all data, we introduce the concept of a "contexton," or unit of context. Contextons are built using "microtypes," which we define as atomic data types representing a simple concept relevant to a domain of interest. Both rely on ontologies, which define a controlled vocabulary for describing a domain of interest. Together, these microtypes and ontologies represent a language that allows users to formally describe all data in that instance in a way that is both *Interoperable* and *Reusable*. 2) Dynamic data types, which make up the vast majority of data, are defined by the users of the system as they are needed, by combining commonly used mathematical data structures with contextons. This "building blocks" approach enables new data types to be defined as needed, with low costs, but also ensures that they are documented in the formal and rigorous manner that is necessary for *Interoperability* and *Reusability* of the data. A limited number of static core types, which are fully specified traditional data structures, are also built using contextons in order to ensure *Interoperability* with the dynamic data. These static core types include the system type *Process*, which is a special core type needed to document the provenance of each data object. 3) All static and dynamic data are referenced in an object graph, where nodes are static or dynamic datasets, and edges are processes. This graph formally annotates the provenance of all data.

In addition to storing ENIGMA data, the Data Clearinghouse includes rich functionality to make the system useful for data analysis, visualization, and managerial oversight. This functionality includes graphing tools, advanced search, an upload wizard, and an API for merging data. ENIGMA data scientists access the Clearinghouse through Jupyter notebooks, running in a shared directory of a server running JupyterHub.

We are also collaborating with the KBase project to harden and deploy these technologies for use by ENIGMA team members as well as all other KBase users. Until the time when these technologies are deployed for general use in KBase, we plan to continue to develop our API and store current ENIGMA datasets using the Data Clearinghouse server. This will ensure that all current ENIGMA data types are compatible with our tools, and that all data can be seamlessly transferred to KBase when our technology is deployed there.

ENIGMA is a Scientific Focus Area Program at Lawrence Berkeley National Laboratory and is supported by the Office of Science, Office of Biological and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.