Laubenbacher (contact PI) 1U01EB024501-01
Model Credibility and Information Management Plan

Project Title: Modular design of multiscale models, with an application to the innate immune response to fungal respiratory pathogens

Project Topic: Development of modular design of multiscale models, clearly separating all dynamic subprocesses. The design is developed around the use case of *Aspergillus fumigatus* infection in the lung and the innate immune response to it. For further information, see https://www.imagwiki.nibib.nih.gov/content/multiscale-modeling-u01-projects

Ten Simple Rules Activities.

Rule Number	**Rule**	**Our activities**
1	Define context clearly	The context is clearly identified in terms of the specific dynamic processes to be modeled explicitly, and the different spatial scales at which each of the processes takes place. The biomedical context focuses on specific aspects of the innate immune response to fungal infections in the lung.
2	Use appropriate data	We are in the process of collecting appropriate data for the models at each of the scales involved. We will also collect a data set from a mouse model that captures the processes at each of the scales simultaneously, important for validation of the integrated multiscale model.
3	Evaluate within context	The model, once complete, will be evaluated within the context of a well-characterized, appropriate animal model, specifically developed by one of the collaborating laboratories for the study of invasive aspergillosis.
4	List limitations explicitly	As the component models are constructed, we explicitly list model assumptions, and other dynamic processes that are relevant, but currently excluded.
5	Use version control	We use a Github repository for version control. We are also developing a more comprehensive platform for data and model management. See below.
6	Document adequately	Documentation is essential and is ongoing. The modular design being developed provides an excellent template for the structure of the documentation.
7	Disseminate broadly	We are planning to disseminate all data and models resulting from the project through a comprehensive information management platform, as well as through public Github repositories and public databases, as appropriate.
8	Get independent reviews	Once a prototype of the model is available, we will involve one or more external evaluators.

9	Test competing implementations	We will implement the model independently through a “conventional” implementation structure, in order to be able to compare and benchmark.
10	Conform to standards	We are conforming to all applicable model and data standards, to the extent these are available. For instance, we are using the ODD protocol for the agent-based model component.

Information Management Activities

The work on this project will generate a range of different types of heterogeneous information, from molecular data to computational models. One of our goals is to develop a comprehensive platform to store, manage this information and make it navigable and useful, leveraging expertise by Kitware Inc., one of the research partners in this project. Since the research partners in this project are geographically dispersed, having such a platform will ensure that the groups can work together effectively, in addition to ease of dissemination to the scientific public. Information includes:

Experimental Data

- Descriptions of in vitro and in vivo experiments performed, essentially the content of a lab notebook, together with images, charts, etc.. This information will be generated by the pulmonary immunology laboratory of Dr. Mehrad, one of the research partners in the project.
- Experimental data of different kinds: molecular, imaging, etc.
- Metadata describing the procedures and algorithms used for generating the data, parameter settings, normalizing procedures, etc.
- Bioinformatics analysis of the data, using a variety of algorithms, resulting in various charts and graphs, and other visualizations. Information includes parameter settings used for algorithms and a list of all choices typically made by a bioinformatician in this setting, as well as procedures for data cleaning etc.

These data will come from the pulmonary immunology laboratory of Dr. Mehrad, one of the research partners in the project, one of the core facilities at the Jackson Laboratory for Genomic Medicine, and the Laubenbacher group.

Simulation Data

- Metadata on model simulations and analysis, including information on algorithms used, assumptions made, model parameters used, random number seeds, etc.
- Metadata related to model analysis, e.g., simulation runs with stats, including parameters such as seeds for random number generators, hardware and software information related to model simulations.
- Visualization data.

These data will be generated by the Laubenbacher group and Kitware.

Models and Methods

- Documentation of various kinds
  - including scientific publications
  - Code

- Mathematical models of various types, discrete, agent-based, PDE, etc., specified in a range of different formats, such as SBML (not so applicable to us), the ODD protocol for agent-based models. These may or may not be equation-based.
- Code for the multiscale model
- Code for component models
- Visualization code
- Data management infrastructure code

This information will come from the Laubenbacher group and Kitware.

Requirements

The information platform should ideally have a variety of capabilities. A user needs to able to navigate this heterogeneous collection of information in various ways. For instance, the user might be looking at a mathematical description of a component model, say an intracellular model for a macrophage in the form of a list of logical rules, and from there

- Move to a simulation of the model with certain external parameter settings (representing particular environmental conditions the macrophage encounters), move to the results for different choices of the parameters (representing a transition of the macrophage to different environmental conditions),
- Look at simulations of the entire multiscale model from the point of view of that macrophage and the two different environmental conditions. This requires a visualization at the level of an alveolar duct.
- Then the user might want to know, for a given model simulation, what portion of the whole lungs exhibit these environmental conditions, for a given set of simulation runs that are archived, and have maybe been used to generate a certain plot that was included in a paper.

Figure 1 contains a schematic description of the platform, followed by a detailed description of requirements.

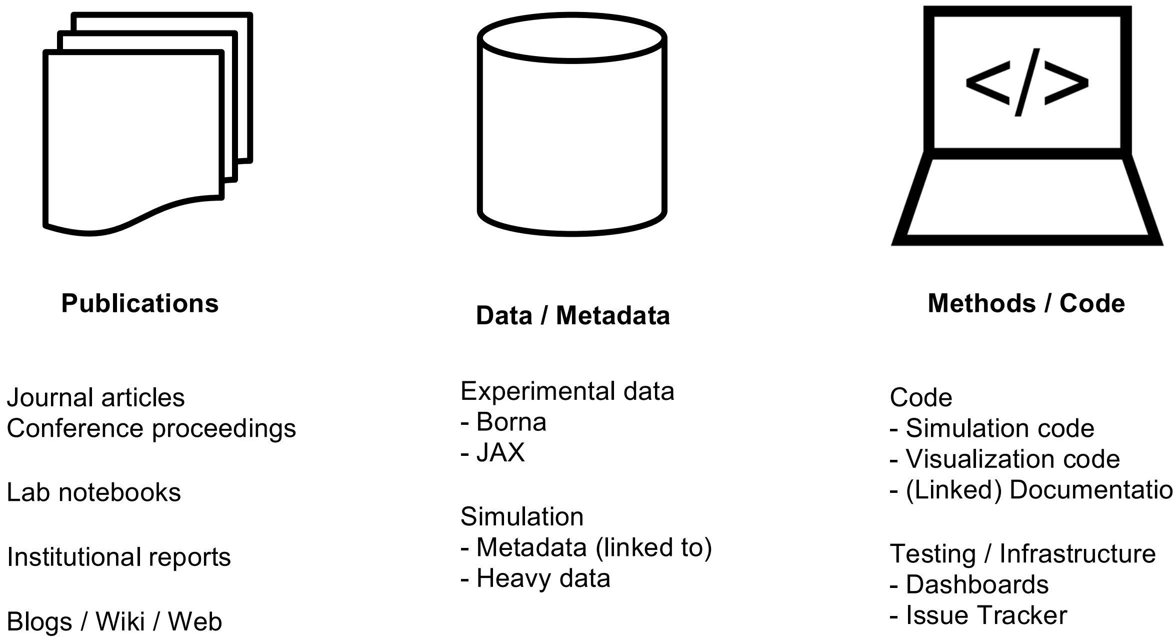

Figure 1. Data types to be integrated

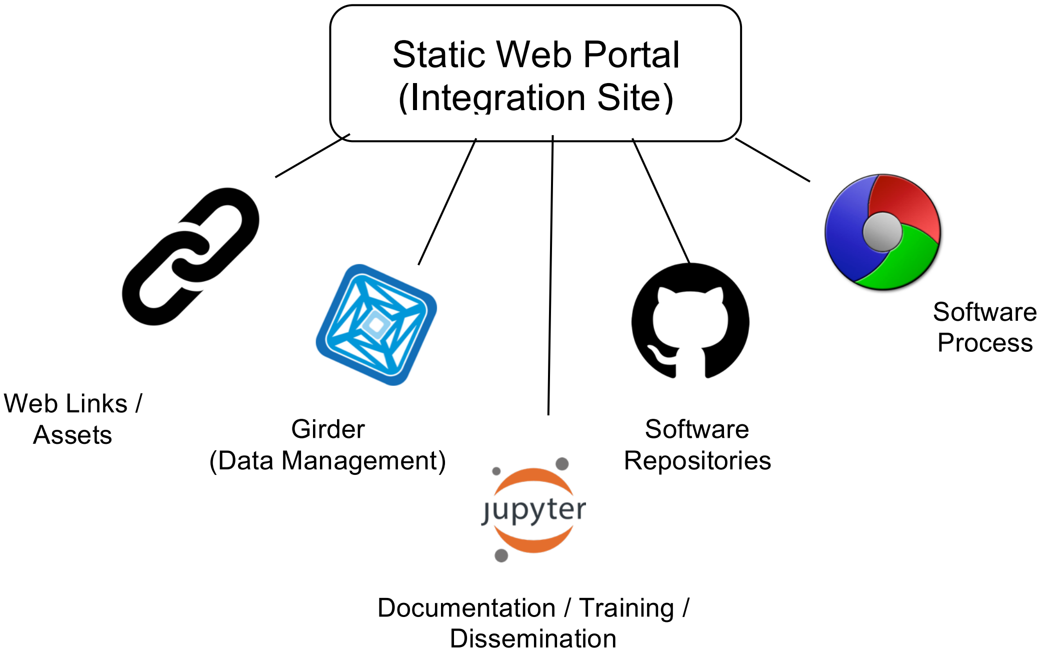

Figure 2. Platform structure.

Our Information Management Plan has been designed to tightly integrate with the Model Credibility plan, addressing the ten rules necessary to ensure model credibility. Additionally this plan facilitates data, software and resource sharing with the overarching goal of supporting Open Science practices. The plan was designed after examining the various forms of data that this U01 project expects to produce and manage (Figure 1).

A critical challenge facing the design of this plan is the extreme heterogeneity of information, ranging from large, heavyweight simulation data, metadata, documentation, publications and even hand notations recorded in a laboratory notebook. Whenever possible standard tools and practices were selected to efficiently implement the plan. Consequently, there is no single tool capable of integrating all of these data in an effective manner. For example, our clinical partner rejected the popular Jupyter Labs notebooks as being too difficult to use in a research lab setting; while github provides an extremely effective process for managing a software process (including version control and testing). Such specialized systems are most effective when used as designed; however integration layers are then required to pull these systems together. The result is a modular, integrated system adopting standard tools with integration pathways expected via Python and JavaScript, with tools like Docker and associated open source repositories for software deployment. Note that this system necessarily supports web access for each sharing and collaboration.

As Figure 2 indicates, there are five major components. 1) Web assets, which consist of items such as static web pages, publications hosted by journals, wikis, and blogs, are linked to their web location. 2) Heavy data and associated metadata is hosted by Girder, a web-based data management platform which supports cloud deployment and managed data analysis. Girder also supports a number of plug-in modules including credentialed access, IO including DICOM, visualization, and provenance. 3) Software, including associated documentation and test data, is hosted in github. This includes both the simulation and visualization software. 4) Longer term, software process is necessary to ensure the stability and reproducibility of computational methods. We will use test-driven development methods with a software quality dashboard tightly integrated with the software repositories. 5) Finally, to support the integration of these components we will use Jupyter Lab to tie together data, computation, and data visualization and analysis. In addition, we will create custom JavaScript applications using VTK.js for visualization and data access demonstrating the ability to create custom analysis applications in cases where Jupyter Notebooks may not be appropriate (or capable). Accessing these components we will build a web portal which will direct users to the appropriate resource(s).

Milestones for Model Credibility and Information Management.

Year 1.

- Develop communication and repository platform
- Set up Github repository

Year 2.

- Continue platform development
- Invite external evaluators
- Upload RNA-seq/protein data, metadata, source code for model architecture, component mathematical models

Year 3.

- Work with evaluators on model assessment/refinement
- Provide open access to all code and data, as appropriate
- Test platform
- Upload in vivo data, metadata

Year 4.

- Make platform publicly available
- Include all data/metadata, source code, mathematical models