Challenges and Future Directions for CIF21 DIBBS: Tripal Gateway, a Platform for Next-Generation Data Analysis and Sharing (Award #1443040).

Stephen Ficklin (PI), Jill Wegrzyn (Co-PI), Frank Feltus (Co-PI), Margaret Staton (Co-PI), Doreen Main (Co-PI), Sook Jung (SP), Kuangching Wang (SP).

Our work as proposed for the NSF DIBBs award #1443040 will provide extended functionality for the Tripal toolkit (http://tripal.info), which is an open-source, freely available software package that provides a framework to assist research groups publish genomic, genetic and related biological data in an online searchable format within Drupal, a popular content-management toolkit. This DIBBs award will allow us to provide greater cyberinfrastructure to these communities by integrating new web services for data exchange between Tripal sites, support for execution of scientific workflows for large-scale datasets, and integration with national research networks (i.e. Internet2).

Throughout our two years of effort on this project we encountered two unexpected challenges. First, because all Tripal sites use the same underlying Chado database schema we hoped to expose data via the web services API following the Chado-style. However, it was clear after speaking with our stakeholders that this was not wanted. They wanted data-specific web services, but, Chado is a large database schema that houses a multitude of data. It would not be practical for us to develop unique web services for each data type. To address this challenge we decided to more fully embrace the semantic web such that each Tripal site administrator could create their own web services API using their own data types. We would ensure exchangeability by redesigning the core software of Tripal to better support controlled vocabularies and use protocols for web service discoverability (e.g. W3C Hydra + JSON-LD). This gives site admins the ability to create their own APIs based on data they house, and sites could exchange data so long as they used common vocabularies. Our redesign of Tripal is almost completed.

Second, when we begin our project we anticipated that Software Defined Networking (SDN) technologies would have matured to a more "production level". However due to differences in vendor implementation of SDN protocols the promise of SDN has not been fully realized. In the interim, we developed the software Big Data Secure Sockets (BDSS) which currently serves the expectation we envisioned for improved data transfer. It uses knowledge of existing network paths, remote repositories transfer protocols, and client software available on the local machine (including parallel file systems) to make smart decisions about data transfer. This software has been released and a publication is in press.

Our final year of work, in 2017, represents our short-term future directions. During this year we will implement the tools we created by integrating three different Tripal-based databases housing genomic data for fruit and forest trees. Additionally we will work with three other legume databases as well. This will represent demonstrable and immediate use of our work. We will also create outreach materials such as tutorials, online documentation, and hold online training sessions to help other Tripal sites adopt these tools. Long term, we hope to create a global interconnected web of federated Tripal sites that have the ability to house and store their own data, following community standards, with the ability to exchange data for large-scale inter-site data analyses.