Supporting data management for structural biologists

West-Life will provide a Virtual Research Environment (VRE) for structural biologists across Europe with users ranging from PhD students to professors. The raw data will be acquired at experimental facilities, and then a series of processing steps will create new data files, leading to the final Protein Data Bank (PDB) file. Larger experimental facilities already have arrangements for storing data, and this is the only possible approach where the technique produces large amounts of data. Smaller facilities will benefit from being able to use EUDAT services.

The Scientific Challenge

The user community consists of a few thousand scientists. Structural biologists used to identify themselves by their preferred technique (as "crystallographers", "electron microscopists" etc). Increasingly, they are targeting larger macromolecular complexes, so research projects now must combine several techniques, at the same time data management and processing are becoming more complex. There is a lot of value in being able to store metadata about the provenance of data ("this file was created by processing those files, using that program with these keywords"). The standard ontology PROV-O expresses most of what this community needs.

Who benefits and how?

This pilot is of benefit to researchers in Structural Biology looking for multiple experimental and computational techniques and visiting multiple experimental facilities/infrastructures to collect their data. The importance of using an integrated approach has been recognized by the foundation of Instruct, an ESFRI Research Infrastructure for integrated structural biology in Europe.

To tackle the complexity issues in structural biology, including spatial relationships on the cellular scale and time resolved structural transitions, structural biologists often need to use complementary techniques in which they are less expert.

There are some technique-specific pipelines that are largely automated for data analysis and/or structure determination, but little is available in terms of automated pipelines to handle integrated datasets. Integrated management of structural biology data from different techniques is lacking altogether. So, enhancing the capabilities of the e-infrastructure available to structural biologists and supporting continued development of computational approaches will be critical in supporting continued growth in their ambition and productivity.

To facilitate both the integration of Structural Biology techniques and the collaborative efforts which will be required to tackle health challenges, there is a strong need for off-the-shelf e-Science solutions providing for high profile projects as well as the long tail of research. These e-Infrastructure solutions for open data sharing, user-friendly access to complex software solutions and computational resources, all of which should be gathered into a virtual research environment, to boost the research output.

Preliminary Results

The single deployment scenario is now implemented (as well as a connection to EUDAT's B2DROP the synchronisation and exchange service for unpublished research data service), while the integration with the West-life portal will be addressed in further releases.

Contacts

- Chris Morris, STFC, chris.morris(a)stfc.ac.uk
- Hans Van Piggelen, SURFsara, hans.vanpiggelen(at)surfsara.nl

Further Information

Scientific Challenges behind the Pilot
Learn more about the EUDAT/West-Life data pilot collaboration here

Services:
B2DROP

Read more