Report from the meeting of the **Interest Group on Structural Biology** at the RDA Third Plenary (Dublin, 27 March 2014)

The Interest Group on Structural Biology (SBIG hereafter) convened at the RDA Third Plenary conference on March 27th, 2014. The agenda and presentations from the two sessions of the meeting are available from the SBIG web page within the RDA site (https://rd-alliance.org/groups/structural-biology-ig/wiki/plenary-3-structural-biology-ig-session.html).
In addition to reviewing various aspects of the state-of-art of data management in structural biology (SB), this meeting provided also an opportunity to establish links with the Interest Group on "Research data needs of the Photon and Neutron Science community", for which biology is a discipline of high relevance.

The following are some key points illustrated in the speakers' presentations on which the attendees formulated a general consensus:

- The complexity of the scientific problems being addressed in SB is steadily increasing. In particular, larger, multi-component (i.e. involving products of multiple genes) biological objects are being tackled more and more. In parallel, the focus is shifting from the macromolecules produced by simpler prokaryotic organisms, useful as model systems, to the macromolecules from higher organisms, i.e. the systems of central relevance for human health
- To cope with the complexity mentioned in the previous point, a clear trend is in place for researchers in SB to use multiple techniques and visit multiple experimental facilities/infrastructures to collect their data. Structural biologists are each expert in one or more techniques, including a deep understanding of data processing and data management facilities. However, they now often need to use complementary techniques in which they are less expert.
- The various individual experimental infrastructures as well as e-infrastructures in the field have developed different solutions to their requirements regarding the management of users, from access/service requests through accounting of instrumentation/CPU usage to interactions occurring after the user's visit/service, such as evaluation of user satisfaction and linking to the resulting publications. Such requirements are largely common but the solutions are not homogenous
- There are some technique-specific pipelines that are largely automated for data analysis and/or structure determination. Little is available in terms of automated pipelines to handle integrated datasets. Integrated management of structural biology data from different techniques is lacking altogether
- Repositories exist for the final structural data (atomic coordinates) as well as, in some cases, for intermediate data generated during the analysis. The provenance and integrity of such data are often an issue as well as their effective compliance to the data generated at the end of the SB determination pipeline
- Journals require deposition of the final structural data in these repositories as a precondition of publication, so compliance is high. However, metadata is often incomplete. The best way to acquire accurate metadata is to integrate data management infrastructure with data processing infrastructure.
- There are no common strategies to address or support the storage of structural biology raw data, after the end of the SB project/grant within which the data were generated. Infrastructures have a policy to maintain the users' data available for only a short period of time after their visit, essentially to allow them to download the data from their home lab. Data volume can sometimes be an issue in this context, because of long data transfer times

Extensive discussion took place at both sessions, involving the general scientific scenario and possible future development of SB as well as practical actions that could be implemented to address (some of) the points above.

A relatively straightforward future improvement that was suggested is the **integration of the data management tools and data warehouses** already in place at the different research infrastructures. This development would take advantage also from the fact that the existing data infrastructures are already designed to deal with distributed/federated experimental facilities. Therefore solutions to some of the issues potentially hindering integration should have been already evaluated or implemented. The resulting integration should (at least initially) aim at providing users with a complete overview of the experiments performed at all the different research infrastructures visited, possibly with a link to the different data storages. Centralization of data storage, even on a temporary basis, did not appear desirable also in view of the data volume potentially involved for each project. The above could be a very useful basis for future initiatives aimed at providing tools for integrated data processing. This proposed activity could be carried out in collaboration with the Photon and Neutron Scattering Interest Group. We thus think that **the above constitutes a ground for this IG to carry on its effort in the form of a Working Group within RDA**.

Experience shows that development of infrastructure is most effective if it is done in close contact with well-chosen pilot research projects that apply the infrastructure. Possible use cases were also extensively discussed in the second session of the meeting. The suggestions that were regarded as being the most feasible (in terms of obtaining access to data) and technically useful are:

- A large multi-component adduct (containing two macromolecular chains or possibly more). A combination of X-ray diffraction, NMR and EM data should be available on the entire adduct and on its components at least partially (that is, for example, each component has been characterized by either X-ray or NMR and the entire adduct by EM and/or another technique). Ideally biophysical measurements and site-directed mutagenesis data should also be available
- A multi-component or multi-domain system for which experimental information on structural as well as dynamic features is available, by a combination of at least two structural techniques