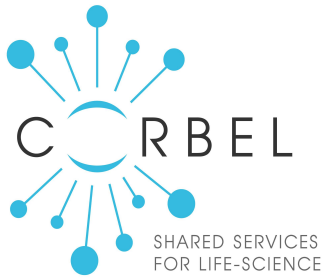

CORBEL WP5 Task 5.3 Briefing Document

Evidence of shared services provided through Instruct structural biology infrastructure to other Biomedical Sciences Research Infrastructures in the CORBEL partnership.

This document forms part of the work undertaken by West-Life (World-wide E-infrastructure for structural biology) Project 675858, of which Instruct is a partner, and is shared with CORBEL (Project 654248) to help establish data sharing processes that can link multi-disciplinary research infrastructures.

Author: Carazo JM, CNB-CSIC (Instruct Centre ES)

West-Life

Project Title:	World-wide E-infrastructure for structural biology	
Project Acronym:	West-Life	
Grant agreement no.:	675858	
Deliverable title:	Publication of a joint document on the usage of structural data in different biomedical Research Infrastructures	
WP No.	3	
Lead Beneficiary:	6 CIRMMP	
WP Title	Networking	
Contractual delivery date:	31 October 2017	
Actual delivery date:	31 October 2017	
WP leader:	Lucia Banci	CIRMMP
Contributing partners:		

Author: Carazo JM, CNB-CSIC

West-Life

Contents

Executive Summary

This document analyses current and emerging usage of structural biology data in the context of different biomedical Research Infrastructures (RIs). It is produced by West-Life as Deliverable 3.5 in close collaboration with other RIs and, specifically, in the context of the European Union CORBEL project (http://www.corbel-project.eu, CORBEL "Coordinated Research Infrastructures Building Enduring Life-science Services").

Structural data, especially in the form of sets of atomic coordinates, are considered by all RIs in the Biomedical arena as very relevant, especially for ELIXIR, Euro-BioImaging and EU-OPENSCREEN. Specific coordination actions among these RIs may provide new synergies to be considered in current and future RI integration projects, such as CORBEL. Another area worth of exploration is the coordination of Instruct-ERIC and other biomedical research infrastructures to fulfill EOSC vision; in this context is particularly important to draw the attention of all biomedical RIs on the coming call of projects under INFRAEOSC-04-2018: Connecting ESFRI infrastructures through Cluster projects.

West-Life

On the nature of Structural biology data

Structural biology data (in short, structural data) refers to detailed information on the shape of macromolecular machines, most of the times reaching atomic or quasi-atomic resolution. The RI in charge of providing access to Structural biology facilities is Instruct-ERIC. The range of technologies available at the different Instruct Centers is large, encompassing from Macromolecular Crystallography (MX), Nuclear Magnetic Resonance (NMR), Small Angle X-ray Scattering (SAXS) to the much newer Direct Electron Detectors-equipped cryo Electron Microscopes (cryo-EM), among many others required for the production and characterization of specimens. In the past, each of the different structural biology technologies has worked in most cases independently, studying different types of specimens or different experimental characteristics. However, the next challenge for structural biology is that the complexity of individual bio-macromolecules and higher order structures cannot be fully addressed by a single experimental approach. Therefore, progress will crucially depend on establishing synergy between several experimental structural techniques.

Typically, the size of the raw data sets being acquired at the different Instruct-ERIC facilities is large (for some of the technologies, the data amount to Terabytes/day), and it is heavily processed both at the facility and at the user laboratory. The end-result of these different layers of processing is the so called “derived data sets”, containing 3D density maps, structural models or interatomic information, together with rich associated metadata. These derived data sets constitute the most relevant piece of information to be considered for “inter Research Infrastructures” cross seeding because of the following characteristics: (1) they are already independent of instrument’s characteristics and facility specificities, (2) they are of a much reduced size (in the order of Gigabytes) than raw data sets, (3), they have attached many “validation/quality” measures that allow users to properly understand the power and limitation of the information they are accessing, thereby already providing a first step towards interpretation and, (4), they are conceptually simpler and, consequently, easier to standardized than raw data.

West-Life

Structural Biology and data sharing

At present, raw data generated within each of the structural techniques, not to mention data generated from methods used by other RIs (e.g. light imaging methods for cells and tissue/organ imaging), are handled using different formats and with different underlying data models that effectively prevent reuse/interoperability at the experimental data level. Although the goal to make raw data interoperable among structural biology techniques is not to be forgotten, it is probably much more practical and feasible in the short/medium term to concentrate on assuring "derived data" interoperability. Indeed, structural biology has been quick to develop and adopt ways to provide effective data sharing, particularly when derived data can be expressed as sets of (x, y, z) atom coordinates (this reduced data representation is often referred as "structural model"). In this way, the most important data base in the field, known as the Protein Data Bank (PDB) (pdb.org), has accepted depositions in the form of atomic coordinates (structural models) for nearly half a century. Data at PDB has traditionally been acquired mainly by X-rays crystallography and NMR, with a minority of cryo-EM data, simply because the latter approach seldom produced atomic or quasi-atomic information. However, the situation has radically changed in the last few years thanks to the combination of several factors, such as: (1) the technological revolution of cryo-EM that has allowed for increased resolution in a large variate of biological systems, (2) key advances in macromolecular crystallography allowing to work with very large macromolecules and with a high degree of automation, and, (3), the advent of "hybrid" methodologies able to effectively combine data from many sources. As a consequence of these changes, several new repositories have been stablished to hold the new data being produced, manly by cryo-EM and integrative models, coordinated under the umbrella of the Worldwide PDB organisation (www.wwpdb.org). Also, new data deposition systems accessing and integrating different technologies have been developed by wwPDB (deposit-1.wwpdb.org). Among these new repositories we highlight the following ones:

For cryo-EM data, EMDB and EMPIAR (www.ebi.ac.uk/pdbe/emdb/; www.ebi.ac.uk/pdbe/emdb/empiar/) archive cryo-EM information. EMDB holds information on "density maps", which are grey scale representations of the Coulomb macromolecular potential at different resolution levels. In the context of this document, density maps are to be understood as derived data, although much less "structured" than atomic models. In turn, EMPIAR address a step forward in the direction of preserving cryo-EM raw data, offering a home for original data sets in the Terabyte range.

As for integrative structural data, PDB-Dev (https://pdb-dev.wwpdb.org/) is a prototype deposition and archiving system for structural models obtained through integrative/hybrid methods.

In terms of Research Infrastructures, and considering that ELIXIR is the RI for Life Science information, the scenario presented in the previous paragraphs clearly recognises the close and long lasting collaboration existing between Instruct-ERIC and ELIXIR.

Regarding specific policies for data management, Instruct-ERIC has had a Data Management Plan (DMP, https://www.structuralbiology.eu/download/do/InstructDataManagementpolicyforUsers.pdf) since 2013, addressing the two different roles of "Facilities" and "Users". This DMP has been at the base of other DMP's for specific projects, such as iNEXT (www.inext-eu.org). The core of this DMP is to drive Structural data even further toward truly sharable principles (technically referred as "FAIR" principles, www.nature.com/articles/sdata201618), aiming at making both raw and derived data easily accessible in a long lasting manner. The existence of these clear principles for data sharing pave the way to an efficient use of Structural data in a very wide context, such as the one involving synergetic work among different RIs, as it will be presented in the following sections.

Synergies among Research Infrastructures (I): The CORBEL project

Probably the best example of concrete synergies achieved by the usage of Structural data at the different RIs can be found in the CORBEL project. CORBEL is a European Union project whose motto already indicates very well the general goal of the initiative; CORBEL: Shared services for Life Science. Eleven Biomedical Research Infrastructures have joined to create a platform aimed at efficiently linking their activities. The role of Structural data in this consortium can objectively be analyzed in the light of the recent CORBEL First Call for Inter RI projects (http://www.corbel-project.eu/1st-open-call.html). A total of 30 eligible projects applied to the call, requesting 80 services and technologies offered by 8 of the 11 RI partners in the project. Out of the 80, Instruct access was solicited 11 times, being one of the top three most demanded RIs.

A deeper analysis of the results of the CORBEL call indicates that Instruct services are required mainly for the two tracks of "Predictive system pharmacology for safer drugs and chemical products" and "Structure function analysis of large macromolecules", but not only there. Indeed, Instruct is also involved in some of the most innovative projects falling in-between CORBEL tracks (the so called cross-Access Track). Regarding those RI that are normally associated with Instruct-ERIC, the two most common partners are Euro-BioImaging and EU-Openscreen, but other RIs are also clearly connected to Instruct-ERIC, such as ELIXIR (for access to Life Science Information), INFRAFRONTIER (for phenotyping of mouse models), EATRIS (for fast track to clinical proof of concept), BBMRI (for access to Biobanks and Biomolecular resources) and ISBE (for access to System Biology approaches). In other words, it is already a fact that in the First Open call for CORBEL projects structural data are being demanded in the context of services provided by seven other RIs. In fact, the single RI that is not part of a CORBEL project in which Instruct-ERIC also participates is EMBRC (The European Marine Biological Resource Center), but this fact may be due to the different level of development of the two infrastructures and, also, to the difficulty to obtain proper samples for Structural biology analysis from marine biology projects.

Synergies among Research Infrastructures (II): West-Life-organized round table on the use of Structural data by RIs

West-life organized a round table in the second iNEXT User Meeting in Brno, June 2017, jointly with iNEXT access project (www.inext-eu.org), to address inter-RI issues (www.structuralbiology.eu/content/bringing-together-the-bio-medical-scientific-communities-the-role-of-research-infrastructures). This event was reported in detail in West-Life Deliverable 3.4. Concentrating on Biomedical Research Infrastructures only, we highlight that representatives of Instruct, Euro-BioImaging, EU-OPENSCREEN and EATRIS attended the event. The clear outcome of this activity from the point of view of use of Structural data was the reiteration of the synergies between these types of data and those coming from cell imaging, especially in the context of drug discovery. The explanation for this situation is simple: a good fraction of structural data are nowadays oriented towards a better understanding of health and disease states, and, specifically, to help in the development of new drugs. Naturally, structural data must be interconnected with other sources of information to be properly put into context and maximally extract their value; to achieve this goal, new data workflows need development to allow the use of datasets from different facilities and technologies; this explains the connection with ELIXIR (for biomedical annotations) and Euro-BioImaging (for functional and live imaging). Of course, functional changes associated with drug testing are of utmost importance, and drug screening is the field of EU-OPENSCREEN.

Still, it should not be forgotten that once the different data sources are appropriately combined, the possibility to propose dynamic models is very appealing (and here is the connection with ISBE), prior to testing in animal models (the role of INFRAFRONTIER), human samples (BBMRI) and, ultimately, clinical trials (EATRIS).

West-Life

Synergies among Research Infrastructures (III): West-Life and the European Open Science Cloud (EOSC)

The European Open Science Cloud (EOSC) represents European Commission view of a virtual environment to store, share and re-use data across scientific disciplines and national border, which should be a reality by 2020. The EOSC will be underpinned by the European Data Infrastructure (EDI). The fits between West-Life and the EOSC is very natural, in that the former produces the structural data that the latter is to assure sharing. In fact, West-Life partners are already involved (or have been involved) in the following EOSC projects:

-MoBrain Competence Center (just ended), within project EGI-Engage

It involved 11 partners, counting with the following West-Life partners: Utrecht University -Coordinator-, CSIC, CIRMMP, INFN and STFC. It addressed topics on cryo-EM cloud deployment, GPU computing and several aspects of virtualization

-WeNMR Thematic Services, within EGI-EUDATA-INDIGO Consortium (starting January 2018)

It involves three West-Life partners: Utrecht University, INFN and CIRMMP. The topic of this project is the support of NMR specific offering of web services.

-Cryo-EM workflows, within EOSC Pilot, as a Science Demonstrator (Jul 2017-June 2018)

Science Demonstrators under EOSC Pilot are rather focused projects, very limited in time (one year) and in budget. It was presented by West-Life partner CSIC with the support of Instruct-Hub and STFC.

Synergies among Research Infrastructures (IV): emerging synergies

Beyond CORBEL and the round table at Brno, we can also highlight some on-going and emerging activities that are opening new possibilities to better address drug design thanks to cross RI interactions, namely “fragment screening” and “new image data bases”.

Regarding fragment screening, this is a key activity in synchrotrons and some NMR facilities, where a library of small molecules (“fragments”) is tested against 3D crystals or soluble proteins already obtained from well-defined pharmacological targets. This process can be very fast and it is crucial for drug discovery. Naturally, there is a clear point in common with EU-OPENSCREEN, the Research Infrastructure organizing and making accessible a large library of currently 140.000 commercial and proprietary compounds collected from European chemists, since these compounds can be the building blocks of many fragment screening approaches. A clear area of collaboration is ready for future work.

Regarding new image data bases, an emerging synergetic activity that remains yet to be further explored is the combination between the very detailed macromolecular information provided by Instruct-ERIC and the functional and live imaging data in which Euro-BioImaging excels. In this latter respect we want to draw the attention to the existing Instruct-EuroBioImaging MoU on image data, as well as the recent and novel integrative data base approaches pioneered by Prof. Swedlow (Willians et al., Nat Methods, November 2016) and Dr. Patwardhan (Iudin et al., Nature Methods, March 2016).

Conclusions

Structural data, especially “derived data”, are of relevance to basically all RIs in the Biomedical arena. Judging by several objective indicators previously introduced in this document, virtually all RIs find structural biology data crucial to be integrated with the specific data produced by each RI. Still, the three RIs that currently have a broader interaction with Structural data and, therefore, also with Instruct-ERIC are: ELIXIR, Euro-BioImaging and EU-OPENSCREEN. Specific coordination actions among these RIs may provide new synergies to be considered in current and future RI integration projects, such as CORBEL. Another area worth of exploration is the coordination of Instruct-ERIC and other biomedical research infrastructures to fulfill EOSC vision; in this context is particularly important to draw the attention of all biomedical RIs on the coming call of projects under INFRAEOSC-*04-2018: Connecting ESFRI infrastructures through Cluster projects*