The Life Cycle of Structural Biology Data

a report of the Structural Biology Interest Group of the Research Data Alliance

Corresponding author: Chris Morris, STFC, Daresbury Laboratory, WA4 4AD

Email: chris.morris@stfc.ac.uk

Contents

Executive Summary

Research data is acquired, interpreted, published, reused, and sometimes eventually discarded. Understanding this life cycle better will help the development of appropriate infrastructural services, ones which make it easier for researchers to preserve, share, and find data.

Structural biology is a discipline within the life sciences, one that investigates the molecular basis of life by discovering and interpreting the shapes and motions of macromolecules. Structural biology has a strong tradition of data sharing, expressed by the founding of the Protein Data Bank (PDB) in 1971 (PDB, 1971; Berman et al 2003). In the early years, data submissions to the archive were made by mailing decks of punched cards. The culture of structural biology is therefore already in line with perspective of the European Commission that data from publicly funded research projects are public data (COM(2011) 882 final).

This report is based on the data life cycle as defined by the UK Data Archive. This is the most clearly defined workflow that the authors are aware of. It identifies six stages: creating data, processing data, analysing data, preserving data, giving access to data, re-using data. Each will be discussed below. However, the data infrastructure for structural biology is not a perfect match for this workflow. For clarity, 'preserving data' and 'giving access to data' are discussed together. We also add a final stage to the life cycle, 'discarding data'.

Changes in research goals and methods have led to some changes in the requirements for IT infrastructure. A common data infrastructure is required, giving a simple user interface and simple programmatic access to scattered data. Progress on these tasks will support the development of workflows that facilitate the use of datasets from different facilities and techniques. The automatic acquisition of metadata can help. Large experimental centres already provide a highly professional data infrastructure. For smaller centres this is onerous - it is desirable that a standard package is provided enabling them to use the European e-infrastructure resources, in a way that integrates with other structural biology resources.

Introduction

In 2015, 9338 new structures were published in the Protein Data Bank, the result of more than 25,000 experimental sessions (see Appendix). Diamond Light Source alone archived more than a petabyte of experimental data during 2015. All these experiments have together a combined data rate greater than that of the Large Hadron Collider.

The physical infrastructure for structural biology includes synchrotrons, which are affordable only by a nation. There are presently 47 in the world (lightsources.org). Each synchrotron provides a number of beamlines for experiments. These usually include some beamlines optimised for macromolecular X-ray crystallography, some for other structural biology techniques including SAXS (Small-Angle X-Ray Scattering) and CD (Circular Dichroism), as well as beamlines for material sciences and other non-biological applications.

A single instrument for NMR (Nuclear Magnetic Resonance) is usually affordable by a university or a company. However, multiple instruments must be used for NMR-based structural biology, because of the need for experiments at different magnetic fields. Thus, typically, investments of the order of 5-10 million euros are required. Because of these rather high costs, a number of large scale facilities have been established around Europe (operating under the former BioNMR and current iNext EU projects) offering the nearly 200 NMR groups in Europe access to very high magnetic fields (Sýkora).

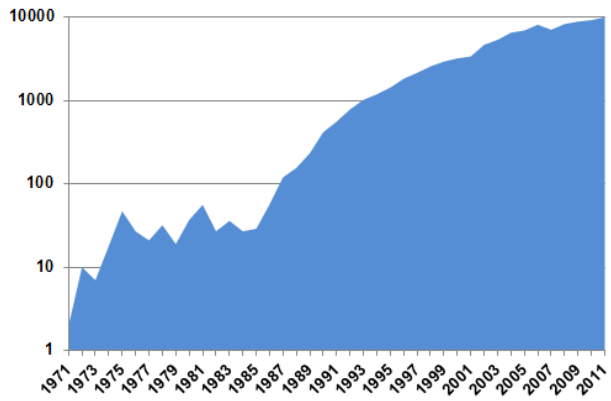


Figure 1. Protein Data Bank: new entries by year (log scale)

Improvements in microscopes, direct electron detectors, and processing software have led to a rapid increase in the number of high resolution cryoEM structures - the “resolution revolution”. This has led in turn to significant investments in electron microscopes around Europe, including dedicated facilities such as NeCEN in Leiden (http://www.necen.nl) and eBIC at Diamond (http://www.diamond.ac.uk/Science/Integrated-facilities/eBIC.html). There is also growing interest in cryoEM from industry, with the formation of the Cambridge Pharmaceutical Cryo-EM Consortium (https://www.mrc.ac.uk/news/browse/pioneering-lmb-research-behind-new-cryo-em-consortium/).

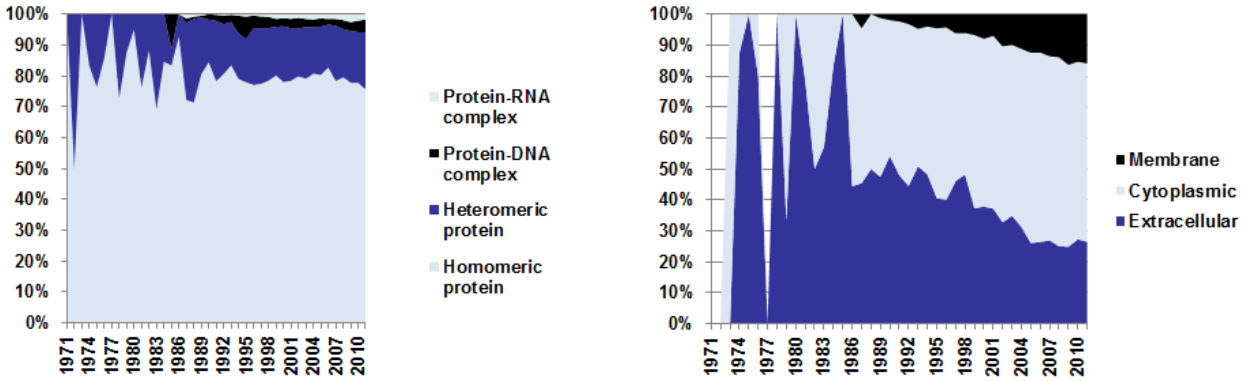


Figure 2. PDB entries grouped by category

Structural biologists are choosing harder targets each year: fewer single proteins, larger macromolecular assemblies, more membrane proteins. Figure 2 shows the increasing proportion of PDB entries belonging to these more difficult categories. Expertise in a single experimental method is not enough to solve these systems.

Berman et al. wrote: “The face of structural biology is changing. Rather than one method being used to determine a single structure, it is becoming more common to use two or more methods and also to study structure at a variety of length scales.” (Berman et al., 2014).

Sali et al. explain “synergy among the input data minimizes the drawbacks of sparse, noisy, and ambiguous data obtained from compositionally and structurally heterogeneous samples. Each individual piece of data may contain relatively little structural information, but by simultaneously fitting a model to all data derived from independent experiments, the uncertainty of the structures that fit the data can be markedly reduced.” (Sali et al.).

A survey of members of Instruct, the ESFRI infrastructure for structural biology, confirmed this picture: 73% were working on eukaryotic rather than prokaryotic systems, and 84% were working on complexes rather that single gene products. As a result, each research team routinely uses 3 or 4 different experimental techniques. However, there are obstacles to this new way of working: 73% say that it is hard to combine software tools for different techniques in integrated workflows (Morris).

The General Life Cycle of Research Data

As Vines et al point out "It is likely that expectations on data sharing will differ between academic communities" (Vines, et al., 2014). This report examines the particular features of the data life cycle within structural biology, and makes recommendations for the next steps in provision of data management facilities.

Sali et al write "The practice of integrative structure determination is iterative, consisting of four stages ...: gathering of data; choosing the representation and encoding of all data within a numerical scoring function consisting of spatial restraints; configurational sampling to identify structural models with good scores; and analyzing the models, ... " (Op. Cit.).There are several descriptions of the life cycle of research data. This essay is based on one of the most cited (UK Data Archive).

It identifies six stages: Creating data, processing data, analysing data, preserving data, giving access to data, re-using data. Each will be discussed below. However, "preserving data" and "giving access to data" are discussed together. We also add a final stage to the life cycle, "discarding data".

Figure 2 shows the life cycle model used in the ICAT software, which manages experimental data at facilities including the ISIS neutron facility and DLS. As will be seen, this facility-centric view is parallel to the project-centric view discussed below.

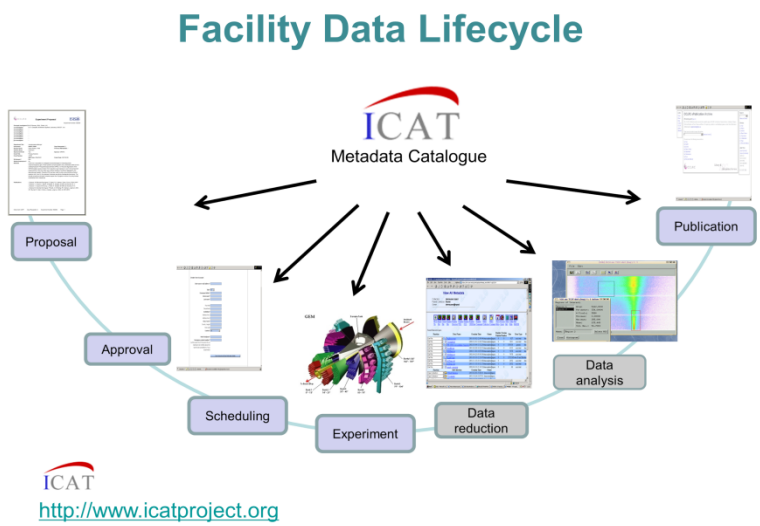
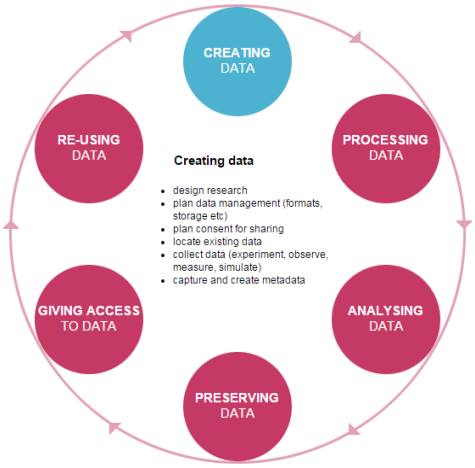


Figure 3

1. **Creating Data**

Primary data is acquired in one or more experiments. Examples include:

- X-ray diffraction at a synchrotron or home source (gigabytes of data)
- NMR spectroscopy (megabytes to gigabytes of data)
- Cryo-electron microscopy (terabytes of data)



This is often referred to as “raw” data. But every observation of nature is mediated by some assumptions, so no data are truly raw. This study uses the term “primary” data, to refer to the first data acquired in a study.

There is metadata describing how the experiment is performed (e.g. the wavelength of the X-rays). Even more important is the provenance of the sample: e.g. how the protein was created and purified. In the case of complexes, this needs to be quite detailed, for example when different components are derived from different species. For NMR experiments, this includes also details of the isotopes used.

The Appendix discusses an estimate that there were more than 25000 experimental sessions in 2015. By March 2015, a total of 3PB of experimental data had been acquired at Diamond (800 million files). This includes all disciplines - about a quarter of experiments there are for the life sciences. This total was up from a reported 1PB a year earlier. The primary data in Single Particle Electron Microscopy is even greater, terabyte images in gray scale.

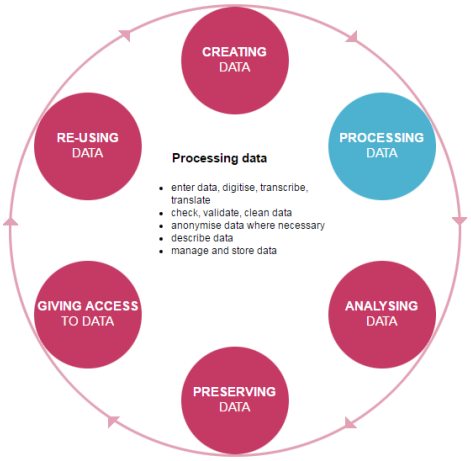
The International Council for Science points out “Publishers have a responsibility to make data available to reviewers during the review process” and that “it is also accessible to ‘post-publication’ peer review, whereby the world decides the importance and place of a piece of research” (Boulton et al.). In line with this, the wwPDB Hybrid/Integrative Methods Task Force recommended: “In addition to archiving the models themselves, all relevant experimental data and metadata as well as experimental and computational protocols should be archived; inclusivity is key.” (Sali et al, op. Cit.). However, there are practical and economic challenges in achieving this, so most experimental facilities expect the users to take their data home with them for processing. Therefore, the responsibility for archiving all relevant data and metadata is left to the individual researcher.

Instruct is the ESFRI infrastructure for structural biology. Its vision is: “We aim to provide strategic leadership for structural biology in Europe by promoting an integrated approach to technology and methodologies. … We provide structural and cell biologists from both industry and academia the opportunity to further their research with cutting-edge technologies sited at Instruct Centres across Europe” (Instruct-vision). In line with the reality discussed above, Instruct’s Data Management Policy says “storage of data is the responsibility of the User to whom it belongs”. However, as the size of datasets increases it becomes impractical for a user to transfer all data to their home institution.

Diamond Light Source (DLS) takes another approach. Like the APS in Chicago, it provides a processing pipeline which often solves the structure without user steering. It also stores the data from synchrotron experiments, and so far has not deleted any experimental data. It also intends to store data from the new electron microscope centre (eBIC). However, it does not issue DOIs for these data. As a result, investigators have to save a copy in another repository if they want to publish the primary data, for example an institutional repository or Zenodo. DLS’s neighbour, the neutron source ISIS, automatically releases primary data with a DOI after three years. An industrial customer of ISIS can pay a fee to keep its data confidential. Similarly at ESRF “The experimental team will have sole access to the data during a three-year embargo period, renewable if necessary. After the embargo, the data will be released under a CC-BY-4 licence” and will be given a DOI [http://www.esrf.eu/datapolicy].

The aim of archiving data at the facility is in line with Instruct’s data management policy, which says: “Instruct Centres are not required to take responsibility for storing data beyond the immediate acquisition visit or the time taken for post experimental analysis if the latter is also provided by the Centre. However, Instruct Centres aspire to offer an archive to store data, especially in cases where the data volume makes this more practical than transferring the data.”.

2. Processing Data: Data Reduction



The first computational processing step typically reduces the data:

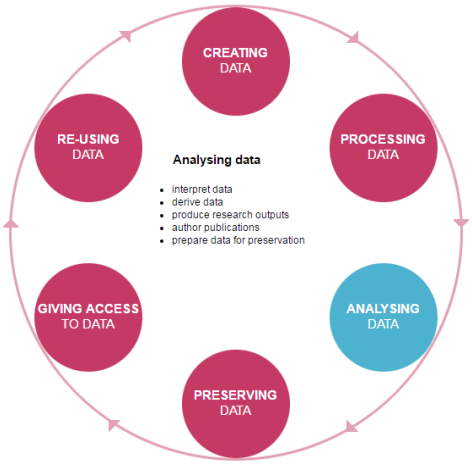
- For Macromolecular X-Ray diffraction (MX), integration of spot intensities and merging of equivalent reflections, reducing the data to megabytes.
- For single particle EM, combining movie frames to make micrographs which reduces the data to hundreds of gigabytes, followed by complex guided workflows to extract particle images and assign them to 2D classes, reducing the data to megabytes.
- For NMR, Fourier transformation actually enlarges the data into gigabytes of processed spectra. This is followed by peak picking and generation of structural restraints.

These procedures give a working dataset, and represent the first stage of interpretation. In MX, one could in principle refine an atomic model against the original diffraction images, making use of the off-reflection diffuse scattering. In NMR, the NMRBox project provides reproducible computing for structure determination [Maciejewski et al].

In cryoEM, one can use the refined model to improve the extraction of a particle from the original micrographs. Thus, the original data has value, and there is a desire to archive it. Nevertheless, most researchers will work with the reduced data, which is simpler to interpret as well as being smaller in size. The complexity of the workflows creates the need for a standard for recording them. Common Workflow Language is a candidate. The accepted standard for data sharing in the community is that the files created in this step should be archived, and should be disclosed when a structure is published. Instruct's Data Management Policy says "supporting data must be deposited in a public database or, in the absence of an appropriate such database, made otherwise available within one year after publication of the results, or within five years after the visit, whichever came first".

3. Analysing Data: Structure Determination and Interpretation

Data reduction is followed by structure determination. For low resolution techniques, the structure may be a volume discretised on a grid or described by an envelope, while for higher resolution the data is interpreted in terms of atomic positions. Sometimes the experimental data is rich enough to determine an approximate structure directly (e.g. by experimental phasing in crystallography), which will later be refined. On other occasions, the "molecular replacement" method involves identifying similar molecules whose structures have already been shared in the PDB, and picking one or more that are a good match for the experimental data as the starting point of refinement.



The refinement process then takes an approximate structure and adjusts it in the light of the experimental data and prior knowledge such as expected stereochemistry (Murshudov et al., 2011). Refinement is an iterative process, which is continued for as long as it continues to produce improvements. Lastly, the structure is subject to a validation step. The PDB provides tools for doing this (Rosato, et al., 2013).

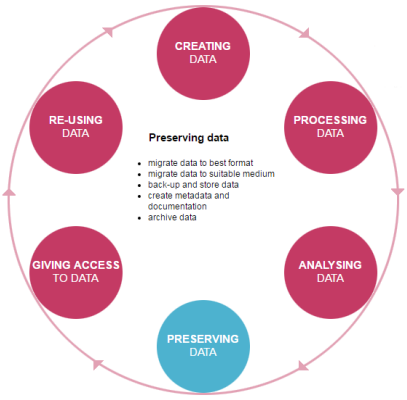
Sali et al. describe this stage as "configurational sampling to identify structural models with good scores; and analyzing the models, including quantifying agreement with input spatial restraints and estimating model uncertainty … all structures are in fact integrative models that have been derived both from experimental measurements involving a physical sample of a biological macromolecule and prior knowledge of the underlying stereochemistry. ".

Some of this processing is performed on scientists laptops and desktops. Some is more computationally intensive but a good match for cloud or grid services (using for example gLite or DIRAC submission mechanisms), notably NMR structure determination and parameter sweeps for more difficult crystallographic problems. The class assignment problem in Electron Microscopy (EM) is a different type of problem, being so intensive in demands for data movement that a high performance cluster is needed, with a good interconnect.

Determining a structure is no longer enough to get published in a high impact journal. The value of structural biology is delivered by interpreting structures, to draw conclusions of wider biological

relevance. Similarly in industry, structures are not determined for their own sake, but as a guide in the development of effective ligands. Hence, the determination and refinement of a structure can be followed by a long period of interpretation. The implications of the structure for known pathways, biochemical results, known effects of mutations, clinical results, etc. need to be worked out. This can lead to a delay in publication, and hence a delay in releasing the structural data.

4. Preserving Data and Giving Access to Data

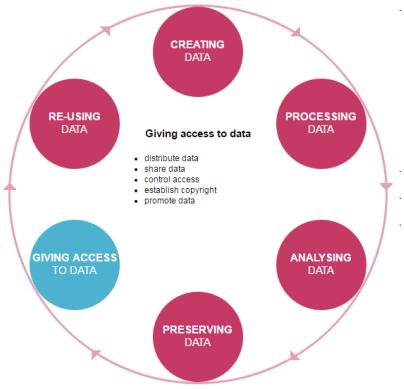


After interpreting the structure, the scientist is ready to write a paper. Journals accept structural papers only if the structure has been shared in the PDB/EMDB. For example the author guidelines for journals published by the International Union of Crystallographers say: "For all structural studies of macromolecules, coordinates and the related experimental data (structure-factor amplitudes/intensities, NMR restraints and/or electron microscopy image reconstructions) must be deposited at a member site of the Worldwide Protein Data Bank" (http://journals.iucr.org/d/services/notesforauthors.html). In practice, at least reduced experimental data is available for 90% of the crystallographic PDB entries, with data missing only for older structures, since it is now mandatory to deposit structure factors for X-ray/Neutron and chemical shifts for NMR experiments. For an archive to be suitable for this use, it must issue DOIs for the deposited datasets.

Given this approach, scientists can and do rely on the PDB/EMDB to preserve not only other people's structures which they wish to see, but also their own.

The PDBx standard (formerly mmCIF, http://mmcif.wwpdb.org/) specifies a rich formal vocabulary for recording experimental conditions and processing methods, including more than 3,000 concepts. But the actual amount of such data recorded in the PDB is disappointing: even crystallogenesis conditions are not reliably reported.

The PDB preserves the refined structural model, and some of the reduced experimental data and sample data, gathered by the data harvesting tool PDB_EXTRACT. However, the larger primary experimental data is not deposited, and other archives have arisen to cater for this need. For all techniques, the Zenodo (https://zenodo.org/) store is available. For X-ray crystallography, diffraction images can be stored using the MyTardis system (http://mytardis.org, Androulakis et al. (2008) doi: 10.1107/S0907444908015540), at https://proteindiffraction.org/ which is provided by the BD2K programme of the NIH, or at the Structural Biology Data Grid (SBgrid) (https://data.sbgrid.org). SBgrid also accepts theoretical models.



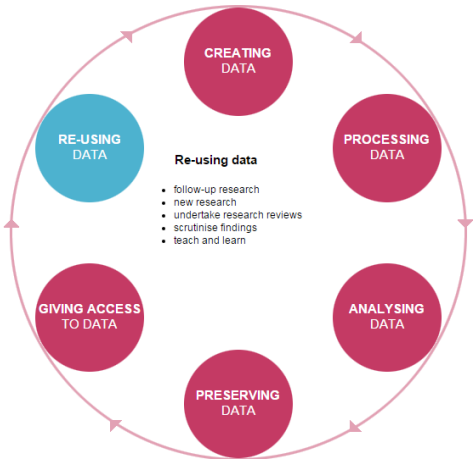
The IUCR points out “For chemical crystallography, IUCr journals require all derived structural models and the processed experimental data sets underpinning them to be submitted for peer review … For macromolecular structures, a validation report is created by database curators when a structural data set is deposited. …. Processed experimental data are also deposited with the structural databases; increasingly reviewers request this (and the raw experimental data) from authors.” (http://www.iucr.org/iucr/open-data). The validation report is not a complete substitute for the diffraction data itself (Minor et al.).

The equivalent recommendations for NMR are presented in (Montelione et al, 2013). NMR data can be archived in the Biological Magnetic Resonance Data Bank (BMRB, http://www.bmrb.wisc.edu, Ulrich et al). This captures more extensive metadata than the PDB does, and some primary data. NMR structural restraints are deposited for all structures. NEF (NMR Exchange Format) is a new common format, developed for representing NMR-derived restraints, and sharing them between structure-generation programs (Gutmanas). This should avoid historical issues with re-interpretation of deposited reduced data.

The EMPIAR service at the EBI will archive raw, 2D electron microscopy images (https://www.ebi.ac.uk/pdbe/emdb/empiar), and the EMDB stores volume maps. EMX is a new metadata format for electron microscopy.

Instruct’s Data Management Policy says “structural data must be either deposited in PDB/EMDB or, as an exception, to be made otherwise available within one year after publication of the results, or within five years after the visit, whichever came first.” In other fields, “Preserving data” and “Giving access to data” are best understood as different stages in the life cycle. In structural biology, both are accomplished by the single step of submission to the PDB/EMDB.

5. Re-using data: Molecular replacement methods and synoptic studies



PDB entries are often reused: in 2012 to 2014 there were 5913 papers citing one or more PDB entries (Bousfield). Instruct's Data Management Policy for Centres says "Instruct intends to provide ways to discover data obtained at the Research Infrastructure, with links to data wherever it was originally collected or processed, and wherever it is currently stored ".

Moreover, "the totality of the data in the PDB provides a rich source of more generalized knowledge about proteins, their molecular biology, and evolution" (Furman et al, 2013). There are more than 500 million downloads per year between all wwPDB partner sites. This is in addition to all the FTP rsync. Many papers are published that report on studies that begin by downloading the whole PDB, then running a program that analyses all the structures to obtain such generalized knowledge. A typical such paper says "First, the UniProt and the PDB database are downloaded from their respective servers, and a local copy of those databases is created." (Baskaran et al).

There are numerous databases and online resources derived from the PDB to facilitate browsing, finding and exploring its entries. These databases contain visualization and analysis tools tailored to specific kinds of molecules and interactions, often also including complex metrics precomputed by experts or external programs, and connections to other non-structural repositories (Abriata). Among the resources provided by West-Life partners, one such database is MetalPDB (http://metalweb.cerm.unifi.it/), which focuses on metal-binding sites in macromolecules (Andreini et al.). Online resources based on the EMDB are also beginning to emerge, for example the PDBeShape volume matching service (http://www.ebi.ac.uk/pdbe/emdb/pdbeshape/) developed as part of the FP7 BioMedBridges initiative. Recent analysis by Monica Sekaran at RCSB shows that "Since 2011, more than 25% of new databases reported by NAR utilized PDB data (119 out of 452 new databases)".

But these are only a part of the reuse. In 2015 there were a total of 534,339,871downloads from the PDB. In the Molecular Replacement method of crystallography, structures from the PDB are used as starting points for the determination of novel structures. Software such as MrBUMP (Keegan and Winn, 2007) and BALBES automates the search of the PDB for suitable structures. Molecular dynamics

simulations (as supported e.g. in BioExcel) reveal the dynamical motion of macromolecules and allow *in silico* experiments, but rely on structures from the PDB for initial conformations.

The PDB-REDO pipeline (Joosten et al., 2014) reuses the reduced data, to repeat the subsequent analysis steps and produce a database of improved structures. In this way, improvements to the refinement software leads to improvements in the results available. The initial effort to populate this database was part of the FP6 project EMBRACE (Pettifer et al.). The West-Life grant has delivered an enhancement to this service to take advantage of a multi-core server. A similar initiative in the field of NMR spectroscopy has been the implementation of the NRG-CING database (Doreleijers et al.). Demonstration of the NRG-CING pipeline on the SURF SARA cloud was achieved in the WeNMR project (Wassenaar et al.). In 2012, a research team in Korea also implemented a database of refined NMR structures, based on statistical potentials (Yang et al.).

Diffraction images stored for example by DLS are reused from time to time, notably by people developing data processing software.

6. Discarding data: obsolete data

At the time of writing, 3,404 PDB entries are marked as obsolete (wwPDB), often by the author and usually because a better sample has been obtained, or a better analysis has been made of the previous data. In rare cases, the erroneous structures were based on fabricated data (Berman, 2010). The PDB now has plans to introduce versioning of structures, so revisions by the author do not break links.

There are examples where self-policing of the structural community, including use of the PDB-REDO server, has been proven effective in detecting incorrect structures of proteins, either during peer review or after publication. This process would be more effective if all datasets were available to reviewers and readers [Kroon-Batenberg et al].

In such cases, the researcher or the institution usually retracts the structure. Unfortunately there have been exceptions. At present, the charter of the PDB only permits it to obsolete an entry if it has been retracted in one of the above ways.

There is also a challenge of incorrect structures for small molecules bound to proteins. Until recently the software tools used did not incorporate prior knowledge of small molecule energetics, and this was not in the expertise of most macromolecular crystallographers either.

Marking a structure as obsolete does not delete the data. Obsoleted coordinates, and the data used to generate them, are valuable to testing new methods of structure quality assessment. For this reason, an archive of annotated obsoleted structures and data should be maintained, separately from the currently recommended model(s). Similarly, data that do not result in a successful structural outcome may have some future value. These data are currently deleted or otherwise lost.

Conclusions: next steps for the data infrastructure for Structural Biology

A report by the International Council for Science points out “Openness and transparency have formed the bedrock on which the progress of science in the modern era has been based. … However, the current storm of data challenges this vital principle through the sheer complexity of making data available in a form that is readily subject to rigorous scrutiny” (Boulton et al).

One of the main obstacles to fully achieve a proper handling of the data life cycle in structural biology is managing the data, which will include datasets acquired in a range of different experimental facilities, some easy to transfer by email or USB stick, and some so large that it is only feasible to process them at source.

A common data infrastructure is required, giving a simple user interface and simple programmatic access to scattered data, so making the facilities offered by EUDAT and INDICO more directly accessible to structural biologists. A significant first step is that Instruct userids are now accepted by EUDAT's B2ACCESS service (https://aarc-project.eu/wp-content/uploads/2016/06/ARIASSO.pdf).

This requirement entails a need for common data formats for the different techniques, and for common data like restraints. Furthermore, there must be tighter and better defined links to the wet lab activities that led to the preparation of the samples used for structural experiments. Although much has been achieved, there is still work to be done to provide full traceability from primers to structure, notably to record construct design, expression conditions, purification conditions, and properties of the sample of soluble protein.

Progress on the above tasks would support the development of workflows that facilitate the use of datasets from different facilities and techniques. In turn, this would lower the barrier for researchers to enter into the field of Integrative Structural Biology, where the complexity of the investigation of the large macromolecular machines of the cell requires an extensive application of multiple structural approaches.

Some data is "orphaned" when the metadata is lost. In the survey of members of Instruct, 26 percent of respondents agreed with the statement "Last year I discarded some samples or files because their provenance was not recorded well enough." As projects get more complicated, this issue becomes worse. This is largely a result of the responsibility for data curation being placed with the individual researcher. The automatic acquisition of metadata would greatly reduce this loss. In particular, by moving data processing to the cloud through the application of largely automated workflow, the acquisition of metadata becomes simple. A further benefit is removing the need for the scientist to perform an extra step of metadata entry.

Large experimental centres already provide a highly professional data infrastructure. For smaller centres this is onerous - it is desirable that a standard package is provided enabling them to use the European e-infrastructure resources, in a way that integrates with other structural biology resources in a seamless manner.

Another obstacle is the burden of installing and using a wide range of software. A crystallographic group will find it very worthwhile to install and keep up to date the CCP4 suite (CCP4). But if a single project uses (for example) AUC, then to find, install, and learn how to use the appropriate software will be burdensome. West-Life will help by cloud provisioning of software and pipelines that apply. In parallel,

protocolized access to software tools via web-based interfaces, as was implemented by the WeNMR project, also provides an efficient approach that allows individual users in any lab worldwide to successfully adopt state-of-the-art tools.

This report will also be presented to a meeting of the RDA Structural Biology Interest Group at the forthcoming RDA Plenary in Barcelona.

Acknowledgements

Thanks to Sameer Velankar, EBI for the data used for graphs of PDB entries. Thanks for discussions to Claudia Alen, Lucia Banci, Alexandre Bonvin, Pablo Conesa, Alfonso Duarte, John Helliwell, Yogesh Gupta, Rob Hooft, John Markley, Brian Matthews, Gaetano Montelione, Antonio Rosato, Sameer Velankar, Matthew Viljoen, Geerten Vuister, John Westbrook, Martyn Winn, and Christine Zardecki. Several of these contributors submitted comments through the mailing list of the RDA Interest Group on Structural Biology, or as speakers at a workshop it held.

Appendix

We estimate that more than 25,000 experimental sessions aimed at structural determination of biological macromolecules are performed each year.

To reach this estimate we first counted the number of new X-ray structures obtained in 2015 at the European Synchrotron Radiation Facility: 633 protein structures were deposited in the PDB in 2015 citing the ESRF as the diffraction source. This is 9% of X-ray structures determined in this period. This corresponds to 1653 experimental sessions, more than 40% of them for macromolecular crystallography (email to author). This suggests that about 16,000 experimental sessions occurred at synchrotrons worldwide.

Next to that, a large number of ligand structures are determined at home sources owned by pharmaceutical companies. These are not usually deposited in the PDB. No estimate can be made here of that activity. The harder, ab initio structures mostly require synchrotron experiments, which is where most academic experiments are conducted.

The Bio-NMR project provided 5610 instrument days over its four-year duration to European scientists studying biological problems by means of NMR spectroscopy. The scientific projects carried out at Bio-NMR facilities resulted in an average of 50-60 structures determined by NMR per year. Many of the larger European universities have their own NMR centers, which are adequately equipped for structural determination of simple soluble proteins or even for the study of protein-protein adducts, and 349 NMR structures were deposited in the PDB in 2015. This suggests that there are about 9000 instrument-days of NMR experiments for structure determination each year, the equivalent of full utilisation of more than 40 magnets.

Adding these 16,000 estimated X-ray sessions and 9000 estimated NMR sessions produces a total of 25,000. An unknown number of sessions for electron microscopy, SAXS, etc., should be added to this.

References

Abriata, 2016. Structural database resources for biological macromolecules. *Brief. Bioinform*. doi:10.1093/bib/bbw049

Andreini et al., 2013. MetalPDB: a database of metal sites in biological macromolecular structures. *NAR*, 41, D312-D319. doi:10.1093/nar/gks1063

Anon., n.d. *Lightsources of the World.* [Accessed 27 June 2016]. Available at: http://www.lightsources.org/regions

Anon., n.d. *Vision.* [Accessed 27 June 2016]. Available at: https://www.structuralbiology.eu/background/instruct/instruct-vision

Baskaran et al, 2014. A PDB-wide, evolution-based assessment of protein-protein interfaces. BMC Structural Biology. 14(22). doi:10.1186/s12900-014-0022-0

Berman et al, 2003. Announcing the worldwide Protein Data Bank *Nature Structural Biology* 10, 980 doi:10.1038/nsb1203-980Berman et al, 2010. Safeguarding the integrity of protein archive. *Nature,* 463(425). doi:10.1038/463425c

Berman et al, 2014. The Protein Data Bank archive as an open data resource. *J Comput Aided Mol Des,* 28(10). doi:10.1007/s10822-014-9770-y

Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L., 2012 . The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure,* Mar , 7(20), p. 391–396. doi: 10.1016/j.str.2012.01.010

Boulton et al, "Open Data in a Big Data World" http://www.icsu.org/science-international/accord/open-data-in-a-big-data-world-long

Bousfield et al, 2016. Patterns of database citation in articles and patents indicate long-term scientific and industry value of biological data resources. *F1000Research,* 5(ELIXIR)(160). doi:10.12688/f1000research.7911.1

CCP4, n.d.[Accessed 28 June 2016] Available at: http://www.ccp4.ac.uk/

Doreleijers et al, 2011. NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *NAR.* doi: 10.1093/nar/gkr1134

Furman et al, 2013. Abstracting knowledge from the Protein Data Bank. *Biopolymers,* 99(3), pp. 183-8. doi:10.1002/bip.22107

Gutmanas et al, 2015. NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *NSMB,* Issue 22, pp. 433-434. doi:10.1038/nsmb.3041

Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A., 2014. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ,* 1(1), pp. 213-220. doi:10.1107/S2052252514009324

Joshua SungWoo Yang et al, 2011. STAP Refinement of the NMR database: a database of 2405 refined solution NMR structures. *NAR,* 40(D1), pp. D525-D530. doi:10.1093/nar/gkr1021

Keegan, R. & Winn, M., 2007. Automated search-model discovery and preparation for structure solution by molecular replacement. *Acta Cryst D,* 63(4), pp. 447-57.

Kroes, N., 2011. Open data: An engine for innovation, growth and transparent governance. [Online] Available at: http://eur-lex.europa.eu/legal-content/EN/AUTO/?uri=COM:2011:0882:FIN

Kroon-Batenburg et al 2017 . Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements. *IUCrJ* 4, pp. 87–99 doi: 10.1107/S2052252516018315

Long et al, 2008, BALBES: a molecular-replacement pipeline. *Acta Cryst D Jan;64(Pt 1):125-32.* doi:10.1107/S0907444907050172

Maciejewski et al. 2017 NMRbox: A Resource for Biomolecular NMR Computation. *Biophys J*. 2017;112(8):1529-34. doi: 10.1016/j.bpj.2017.03.011.

Montelione, G. T. et al, 2013. Recommendations of the wwPDB NMR Validation Task Force. Structure 21(9), pp. 1563-1570. doi:10.1016/j.str.2013.07.021

Minor, Dauter, Helliwell, Jaskolski, and Wlodawer, "Safeguarding structural data repositories against bad apples", Structure 24(2) p216-220, 2016, doi: 10.1016/j.str.2015.12.010

Morris, n.d. *Software and Data Management Tools for Integrated Structural Biology.* [Accessed 28 June 2016]. Available at: https://www.structuralbiology.eu/update/download/do/Instruct_Software_Survey.pdf

Murshudov, G. N. e. a., 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta. Cryst. D,* 67(4), pp. 355-367. doi:10.1107/S0907444911001314

PDB, 1971. Protein Data Bank. *Nature New Biology.*

Pettifer et al, 2010. The EMBRACE web service collection. *NAR,* Volume 38, pp. 683-8. doi:10.1093/nar/gkq297

RCSB, n.d. *Obsoleted PDB Entries.* [Accessed 28 June 2016] Available at: http://www.rcsb.org/pdb/home/obs.do

Rosato, A., Tejero, R. & Montelione, G. T., 2013. Quality assessment of protein NMR structures. *Curr Opin Struct Biol,* 23(5), pp. 715-24. doi:10.1016/j.sbi.2013.08.005

Sýkora, n.d. *NMR, MRI, ESR and NQR Centres and Groups* [Accessed 28 June 2016] Available at: http://www.ebyte.it/library/NmrMriGroups.html

Sali et al , 2015. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure,* 7(23), pp. 1156-67. doi:10.1016/j.str.2015.05.013

Ulrich et al, 2008 BioMagResBank. *Nucleic Acids Res.* 36(Database issue):D402-8. PubMed PMID: 17984079.

UK Data Archive, n.d. *Research Data Lifecyle.* [Accessed 27 June 2016] Available at: http://www.data-archive.ac.uk/create-manage/life-cycle

Vines, T. H. et al., 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology,* 24(1), pp. 94-97.

Wassenaar et al, 2012. WeNMR: Structural Biology on the Grid. *J. Grid. Comp.,* Issue 10, pp. 743-767.

Yang et al, 2013. STAP Refinement of the NMR database: a database of 2405 refined solution NMR structures.. *NAR,* 40(Database issue), pp. D525-30.

Glossary

Eukaryote. A higher organism, with a distinct nucleus in the cell, e.g. human or yeast. As contrasted with a **prokaryote**. Protein expression is more elaborate in eukaryotes, with more folding mechanisms and more extensive post-translational modification.

Heteromeric. A complex containing more than type of molecule.

Homomeric. A complex consisting of several molecules of the same type, e.g. P53 functions as a unit containing four identical protein molecules.

Instruct. The ESFRI infrastructure for structural biology.

Metadata. Data about data, e.g. provenance.

MX. Macro-molecular X-ray Diffration.

Prokaryote. A bacterium, e.g. E. coli. As contrasted with a **eukaryote**.

SAXS. Small-angle X-ray Scattering.

X-ray Diffraction. An experimental technique that exposes a crystal to a beam of X-rays, producing a “diffraction pattern”, from which the structure of the contents of the crystal can be determined.