
PRACTICE PAPER

The Life Cycle of Structural Biology Data

Chris Morris

STFC, Daresbury Laboratory, WA4 4AD, GB
chris.morris@stfc.ac.uk

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. The culture of structural biology is therefore already in line with the perspective that data from publicly funded research projects are public data.

This review is based on the data life cycle as defined by the UK Data Archive. It identifies six stages: creating data, processing data, analysing data, preserving data, giving access to data, and re-using data. For clarity, ‘preserving data’ and ‘giving access to data’ are discussed together. A final stage to the life cycle, ‘discarding data’, is also discussed.

The review concludes with recommendations for future improvements to the IT infrastructure for structural biology.

Keywords: Structural biology; virtual research environment; data life cycle; open access; open science

Introduction

In 2016, 24408 new structures were released in the Protein Data Bank (PDB) (**Figure 1**). Diamond Light Source (DLS) alone acquired about two petabytes that year. 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, presently 47 in the world (Lightsources n.d.). Each synchrotron provides a number of beamlines for experiments, including some beamlines optimised for macromolecular X-ray crystallography, some for other structural biology techniques including Small-Angle X-Ray Scattering and Circular Dichroism, as well as beamlines for non-biological applications.

Structural biologists who use Nuclear Magnetic Resonance (NMR) need experiments at different magnetic fields. Thus investments of the order of 5–10 million euros are required. A number of large scale facilities have been established around Europe (operating under the former BioNMR and current iNext EU projects) supporting nearly 200 NMR groups in Europe (Sýkora n.d.).

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 (NeCEN) and eBIC at Diamond (eBIC). There is also growing interest in cryoEM from industry, with the formation of the Cambridge Pharmaceutical Cryo-EM Consortium.

Structural biologists are choosing harder targets each year: fewer single proteins, more membrane proteins, and more eukaryotic 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. A survey of members of Instruct, the European infrastructure for structural biology, confirmed this picture: 73% were working on eukaryotic rather than prokaryotic systems, and 84% were working on

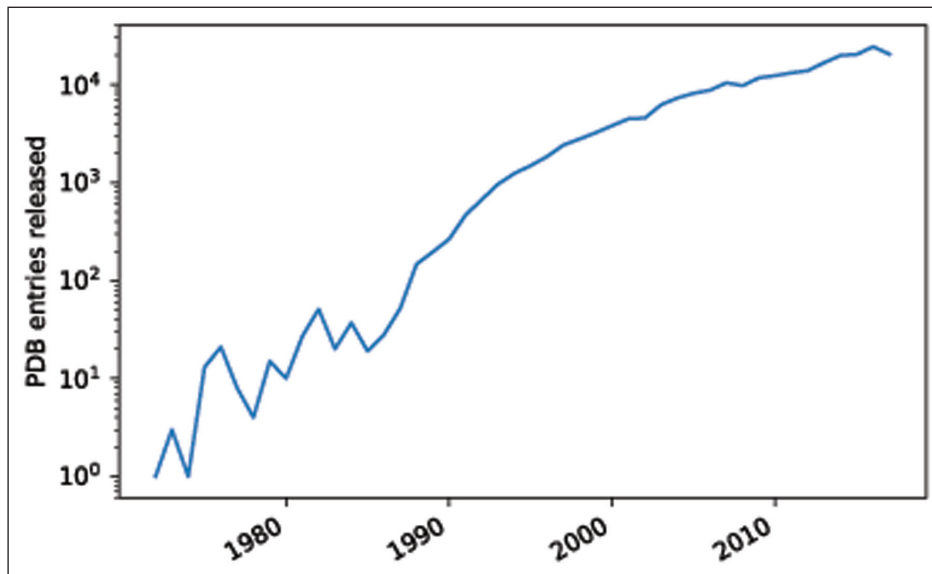


Figure 1: Protein Data Bank: new entries by year (log scale).

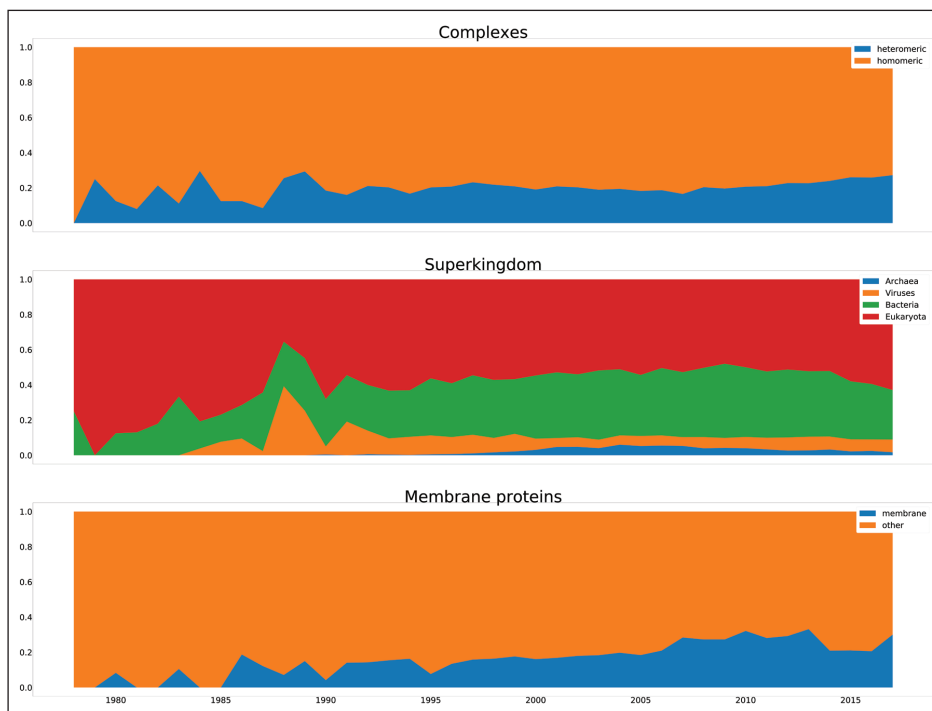


Figure 2: PDB entries grouped by category.

complexes rather than single gene products. 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 n.d.).

There are several descriptions of the life cycle of research data. This essay is based on one of the most cited (UK Data Archive n.d.). It identifies six stages: creating data, processing data, analysing data, preserving data, giving access to data, and re-using data. Each will be discussed below. However, 'preserving data' and 'giving access to data' are discussed together. A final stage is also discussed, 'discarding data'.

The methodology of this review was focussed on the question of what improvements in the IT infrastructure would most benefit for structural biology, and the review concludes with recommendations on this topic. The methods of investigation were a literature review, collection of data from public databases, and a survey of structural biologists.

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. This study uses the term 'primary' data, to refer to the first data acquired in a study.

Metadata describes 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. For NMR experiments, this includes also details of the isotopes used.

Nearly 13PB of experimental data had been stored at DLS by the end of 2017. This includes all disciplines – about a quarter of experiments there are for the life sciences. The primary data in Single Particle Electron Microscopy is even greater, terabyte movies in grey 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. n.d.). In line with this, the Hybrid/Integrative Methods Task Force of the world-wide PDB (wwPDB) recommended: 'In addition to archiving the models themselves, all relevant experimental data and metadata as well as experimental and computational protocols should be archived' (Sali et al. 2015). 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.

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 also becomes impractical for a user to transfer all data to their home institution.

DLS takes another approach. Like the Advanced Photon Source 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 (ESRF Data Policy).

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 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. In NMR, the NMRBox project provides reproducible computing for structure determination (Maciejewski et al. 2017).

Analysing Data: Structure Determination and Interpretation

Data reduction is followed by structure determination. 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. Alternatively 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 typical stereochemistry (Murshudov et al. 2011). This is continued for as long as it continues to produce improvements. Lastly, the structure is validated. The PDB provides tools for doing this (Rosato et al. 2013). Sali concludes: '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 ...' (Sali et al. 2015).

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 (IUCR) 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' (IUCR Notes for Authors). In practice, at least reduced experimental data is available for 90% of the crystallographic PDB entries, with data missing only for older structures. Scientists rely on the PDB/EMDB to preserve not only other people's structures which they wish to see, but also their own.

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. The PDBx standard (mmCIF), 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 crystallography conditions are not reliably reported.

The larger primary experimental data is not deposited in the PDB, so other archives have arisen to cater for this need. For all techniques, the Zenodo store is available. For X-ray crystallography, diffraction images can be stored using the MyTardis system (Androulakis et al. 2008), at <https://proteindiffraction.org/> which is provided by the BD2K programme of the NIH, or at the Structural Biology Data Grid (SBgrid). 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." (IUCR Open Data). The validation report is not a complete substitute for the diffraction data itself (Minor et al. 2106).

The equivalent recommendations for NMR are presented in Montelione et al, 2013. NMR data can be archived in the Biological Magnetic Resonance Data Bank (Ulrich et al. 2008). 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 et al. 2015). 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, and the EMDB stores volume maps. EMX is a new metadata format for electron microscopy.

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.

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 et al. 2016). There are more than 500 million downloads per year between all wwPDB partner sites. 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. 2014).

These are only a part of the reuse. In 2017 there were a total of 679,421,200 downloads 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 (Long et al. 2008) 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. They 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. The NRG-CING database (Doreleijers et al. 2011) is a similar initiative for NMR.

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

Discarding Data: Obsolete Data

At the time of writing, 3,404 PDB entries are marked as obsolete (RCSB), usually because a better sample has been obtained or a better analysis has been made of the previous data. 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-Batenburg et al. 2017).

In such cases, the researcher or the institution usually retracts the structure. Unfortunately there have been exceptions. In rare cases, the erroneous structures were based on fabricated data (Berman et al. 2010). 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. 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. n.d.).

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 e-Infrastructures more directly accessible to structural biologists.

Combined methods demand 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. There is still work to be done to provide full traceability from gene to structure, notably to record construct design, expression conditions, purification conditions, and properties of the sample of soluble protein.

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 workflows, the acquisition of metadata becomes simpler.

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. Cloud provisioning of software and pipelines can help.

Acknowledgements

Thanks to John Helliwell, Manchester for discussions including the point that reviewers must have access to primary data. 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, Nurul Nadzirin, 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.

This work has been done as part of the WestLife VRE (www.west-life.eu), a project funded by the European Commission contract H2020-EINFRA-2015-1-675858.

Competing Interests

The author has no competing interests to declare.

Author Information

Chris Morris is a data analyst and software project manager at the Daresbury Laboratory, STFC. He is the project manager for West-Life, a Horizon 2020 project developing a Virtual Research Environment for structural biology. He is also involved in ADDoPT, an Innovate UK grant to support the design of pharmaceutical dosage forms. He has been a software developer for twenty five years. He says 'Eventually I realised that the coding is not the hardest part of the job'.

References


- Androulakis, S**, et al. 2008. Federated repositories of X-ray diffraction images. *Acta Cryst*, D64: 810–814. DOI: <https://doi.org/10.1107/S0907444908015540>
- Baskaran, K**, et al. 2014. A PDB-wide, evolution-based assessment of protein-protein interfaces. *BMC Structural Biology*, 14(22). DOI: <https://doi.org/10.1186/s12900-014-0022-0>
- Berman, HM**, et al. 2010. Safeguarding the integrity of protein archive. *Nature*, 463(425). DOI: <https://doi.org/10.1038/463425c>
- Boulton**, et al. 2015. Open Data in a Big Data World. https://www.icsu.org/cms/2017/04/open-data-in-big-data-world_long.pdf [Last accessed 22 June 2018].
- Bousfield, D**, 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: <https://doi.org/10.12688/f1000research.7911.1>
- CCP4**. n.d. Available at: <http://www.ccp4.ac.uk/> [Last accessed 28 June 2016].
- Doreleijers, JF**, et al. 2011. NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB. *NAR*. DOI: <https://doi.org/10.1093/nar/gkr1134>
- eBIC**. n.d. <http://www.diamond.ac.uk/Science/Integrated-facilities/eBIC.html> [Last accessed 28 June 2016].
- ESRF Data Policy. n.d. <http://www.esrf.eu/datapolicy> [Last accessed 27 June 2016].
- Gutmanas, A**, et al. 2015. NMR Exchange Format: a unified and open standard for representation of NMR restraint data. *NSMB*, 22: 433–434. DOI: <https://doi.org/10.1038/nsmb.3041>
- IUCR**. 2015. Open Data. <http://www.iucr.org/iucr/open-data> [Last accessed 27 June 2016].
- IUCR**. Notes for Authors. n.d. <http://journals.iucr.org/d/services/notesforauthors.html> [Last accessed 27 June 2016].
- Joosten, RP, Long, F, Murshudov, GN and Perrakis, A**. 2014. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ*, 1(1): 213–220. DOI: <https://doi.org/10.1107/S2052252514009324>
- Keegan, R and Winn, M**. 2007. Automated search-model discovery and preparation for structure solution by molecular replacement. *Acta Cryst D*, 63(4): 447–57. DOI: <https://doi.org/10.1107/S0907444907002661>

- Kroon-Batenburg, LMJ**, et al. 2017. Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements. *IUCrJ*, 4: 87–99. DOI: <https://doi.org/10.1107/S2052252516018315>
- Lightsources**. *Lightsources of the World*. Available at: <http://www.lightsources.org/regions> [Last accessed 27 June 2016].
- Long, F**, et al. 2008. BALBES: a molecular-replacement pipeline. *Acta Cryst D Jan*, 64(Pt 1): 125–32. DOI: <https://doi.org/10.1107/S0907444907050172>
- Maciejewski, MW**, et al. 2017. NMRbox: A Resource for Biomolecular NMR Computation. *Biophys J*, 112(8): 1529–34. DOI: <https://doi.org/10.1016/j.bpj.2017.03.011>
- Montelione, GT**, et al. 2013. Recommendations of the wwPDB NMR Validation Task Force. *Structure*, 21(9): 1563–1570. DOI: <https://doi.org/10.1016/j.str.2013.07.021>
- Minor, W, Dauter, Z, Helliwell, JR, Jaskolski, M and Wlodawer, A**. 2016. Safeguarding structural data repositories against bad apples. *Structure*, 24(2): 216–220. DOI: <https://doi.org/10.1016/j.str.2015.12.010>
- mmCIF**. n.d. <http://mmcif.wwpdb.org/> [Last accessed 27 June 2016].
- Morris**. n.d. *Software and Data Management Tools for Integrated Structural Biology*. Available at: https://www.structuralbiology.eu/update/download/do/Instruct_Software_Survey.pdf [Last accessed 28 June 2016].
- Murshudov, GN**, et al. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Cryst D*, 67(4): 355–367. DOI: <https://doi.org/10.1107/S0907444911001314>
- NeCEN**. n.d. <http://www.necen.nl> [Last accessed 27 June 2016].
- RCSB**. n.d. *Obsoleted PDB Entries*. [Last accessed 28 June 2016] Available at: <http://www.rcsb.org/pdb/home/obs.do>.
- Rosato, A, Tejero, R and Montelione, GT**. 2013. Quality assessment of protein NMR structures. *Curr Opin Struct Biol*, 23(5): 715–24. DOI: <https://doi.org/10.1016/j.sbi.2013.08.005>
- Sali, A**, et al. 2015. Outcome of the First wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure*, 23(23): 1156–67. DOI: <https://doi.org/10.1016/j.str.2015.05.013>
- Sýkora**. n.d. *NMR, MRI, ESR and NQR Centres and Groups* [Last accessed 28 June 2016] Available at: <http://www.ebyte.it/library/NmrMriGroups.html>.
- Ulrich, EL**, et al. 2008. BioMagResBank. *Nucleic Acids Res*. 36(Database issue): D402–8. PubMed PMID: 17984079.
- UK Data Archive**. n.d. *Research Data Lifecyle*. [Last accessed 27 June 2016] Available at: <http://www.data-archive.ac.uk/create-manage/life-cycle>.

How to cite this article: Morris, C. 2018. The Life Cycle of Structural Biology Data. *Data Science Journal*, 17: 26, pp. 1–7, DOI: <https://doi.org/10.5334/dsj-2018-026>

Submitted: 23 August 2017 **Accepted:** 01 October 2018 **Published:** 12 October 2018

Copyright: © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 