Institut Max Von Laue – Paul Langevin (ILL)

Scientific Data Policy

July 2017

Motivation: Central facilities for neutron scattering and synchrotron x-rays in Europe are working together increasingly to develop and share infrastructure for the data collected there. Such co-operation should make it easier and more efficient for users to access and process their data, and provide more secure means of storage and retrieval. It should also increase the scientific value of the data by opening it up to a wider community for further analysis and fostering new collaborations between scientific groups. Ultimately this should improve the quality and quantity of publications arising from such data. However, with these developments comes a need to define how such data and associated results are stored and made accessible, and for this a common data policy structure has been defined by neutron scattering and synchrotron x-ray facilities to provide a suitable working framework. The ILL data policy respects these measures.

1. General principles

1.1 This scientific data policy pertains to the ownership of, the curation of and access to ILL data, meaning experimental raw data and metadata collected and/or stored at the ILL, as well as any processed data and results that are generated and/or stored at the ILL.

1.2 Acceptance of this data policy is a condition for the award of ILL beamtime.

1.3 ILL users must not attempt to access, exploit or distribute any ILL data unless they are entitled to do so under the terms of this data policy.

1.4 Deliberate infringements of this data policy may lead to denial of access to such data and/or denial of future beamtime requests at the ILL.

1.5 All ILL data will be subject to the data protection legislation of France, being the country in which the data are stored.

1.6 The unique identifier for an ILL data set is the DOI (Digital Object Identifier), which incorporates the ILL beamtime proposal number and, when associated with the name of the ILL instrument(s) used, refers to a specific ILL experiment.

2. Definitions

For the purposes of this data policy:

2.1 The term **raw data** pertains to data collected directly from ILL instruments during experiments. This definition includes data that are created either automatically by instrument-control software (e.g. detector counts, angles, time stamps, etc) or manually by ILL staff and/or the experimental team, but which have not yet been reduced or processed by any data-treatment software. In practice, raw data are curated in the /rawdata/ subdirectory corresponding to the performed experiment.

2.2 The term **metadata** describes contextual information generated at the ILL that is complementary to raw data and possibly useful for subsequent data treatment. Metadata include (but are not limited to) the ILL beamtime proposal (including abstract but not the 2-page scientific description), logfiles and parameter surveys generated by the instrument-control software, the instrument configuration (e.g. wavelength), the sample environment, the sample description and state points (e.g. temperature and pressure), the content of the instrument notebook, and other logistical information. In practice, electronic metadata are curated in the ILL proposal database as well as in the /histo/ and /logfiles/ (and usually also /rawdata/ and /processed/) subdirectories corresponding to the ILL experiment.

2.3 The term **processed data** refers to raw data and/or metadata that have been processed or reduced by data-treatment software and then curated in the /processed/ subdirectory corresponding to the performed experiment. As such, processed data can be also considered to be results as defined below.

2.4 The term **data** without qualifiers refers to the ensemble of raw, meta- and processed data.

2.5 The term **results** refers to treated/analysed data and/or the associated scientific conclusions, intellectual property, and other outcomes, resulting from ILL experiments, and stored either at or outside of the ILL. This does not include publications, but can include files that are stored in the /processed/ subdirectory (and within its subdirectories) corresponding to the performed experiment.

2.6 The term **curation** denotes the long-term storage, backup and protection of data and results in a manner that respects and guarantees the prescribed access-rights.

2.7 The term **long-term** means a minimum of 5 years and the ILL will thrive for at least 10 years. Obviously, the exact time-scale for long-term data curation may depend on the type and volume of data stored as well as financial limitations. The ILL therefore reserves the right to restrict the curation periods in consultation with the respective communities for high data-rate instruments.

2.8 By default, the term **access** refers to read-access only in the case of raw data, to read-access and possibly also write-access for metadata, and to read+write-access for processed data and results. As write-access to processed data curated at the ILL is tantamount to performing some level of data treatment/analysis, there is no need to specify separate access rights for data treatment as opposed to data retrieval by the authorized members (definition below) of a beamtime proposal.

2.9 An **Access Control List (ACL)** is a software tool that lists and attributes rights to users for accessing data files (here: read+write, read-only, or no access). The ILL's ACLs operate directly on the experimental data repository, thus controlling all access from all ILL workstations, whether running Windows/PC, Macintosh or Linux.

2.10 The term **open access** (not to be confused with “public domain”) means freely available and useable by the community at large, i.e. being unprotected by copyright or patent and subject to use by anyone respecting a few basic requirements. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based. The ILL provides open access via the ILL Data Portal (data.ill.eu) in the form of download access only (i.e. read-only access) and subject to certain conditions and obligations.

2.11 Data enjoying open access are called **open data**, which for the ILL refers to data that have been released under the terms of the license CC-BY (http://creativecommons.org/licenses/) that obliges the users of such data to cite the ILL and the corresponding experiments and scientific teams as the source of the data. Therefore, the ILL's open data does *not* reside in the public domain.

2.12 The term **public research** refers to research done through peer review (e.g. the ILL beamtime proposal system) and intended for publication.

2.13 The term **proprietary research** refers to research done through purchased (commercial) access to the ILL (beamtime and associated facilities), and there is no obligation to publish the data or results.

2.14 The term **principle investigator** (PI) pertains to the main proposer identified on the ILL beamtime proposal. For experiments outside of the ILL proposal system, the PI is the person initiating or performing the experiment.

2.15 The PI and the ILL local contact(s) for a given proposal are by default **proposal managers**, meaning that they can modify the rights of others to access data (raw, meta- or processed) pertaining to that proposal.

2.16 The term **proposal team** includes the main proposer (also referred to as the PI - above) and all co-proposers of the ILL beamtime proposal.

2.17 The term **experimental team** includes all participants, including the ILL local contact(s), in the experiment(s) performed for a given proposal. Note that a given proposal can implicate several instruments and/or several beamtime scheduling periods.

2.18 An **authorized member** of a proposal refers to a person having full access to the data (raw, meta- and processed) for that proposal. The authorized members include by default all persons in the proposal team as well as the experimental team(s) for the experiment(s) of that proposal, and therefore includes the proposal managers who have the power to add additional authorized members.

2.19 The term **on-line catalogue** designates a computer database and accompanying software tools allowing access to ILL-curated data (raw, meta- and processed) from both inside and outside the ILL. For example, the **ILL Data Portal** (data.ill.eu, aliased to data.ill.fr) is a web-based on-line catalogue that allows access to data that are indexed by the proposal number and by the DOI as unique identifer. Such data are curated in the ILL proposal database as well as in the proposal's file directory that is subdivided into /rawdata/, /logfiles/, /histo/ and /processed/ subdirectories.

2.20 **A registered user** refers to any individual who has been granted an ILL User Club account, which can be requested via the login webpage for the ILL User Club (userclub.ill.eu).

3. Policy towards data pertaining to ILL experiments

3.1 Ownership of data

3.1.1 All data (raw data, metadata and processed data) obtained via public research conducted at the ILL are destined for open access (i.e. to become open data) after an embargo period of 3 to 5 years after the experiment, with the ILL acting as the curator during and after the embargo period.

3.1.2 All data obtained via proprietary research conducted at the ILL will be owned exclusively by the client who purchased the access. Proprietary users must agree with the facility management how they wish their data to be managed before the start of any experiment.

3.1.3 Data from CRG (Collaborating Research Group) instruments obtained during ILL beamtime are subject to the same rules as data from ILL instruments.

3.1.4 Data from CRG instruments that are obtained during CRG beamtime, but for which the CRG has chosen to make use of the ILL's software tools and IT infrastructure for data curation (i.e. data storage, protection and access) are subject to the same rules as data from ILL instruments or ILL beamtime, including that the data will become released as open data after an embargo period (see below). Otherwise data obtained during CRG beamtime is not considered to be curated by the ILL, so that the ownership and curation of such data is to be managed by the CRG. Naturally, the ILL and the CRG have the possibility to negociate special arrangements that are to their mutual benefit.

3.2 Curation of data by the ILL

3.2.1 All raw data will be curated within files of well-defined format (so-called “numor” files) in the /rawdata/ subdirectory corresponding to the performed experiment. The ILL will strive to make available the means of reading these files (sometimes in a very special format) over the long-term.

3.2.2 Metadata that are automatically captured by instruments will be curated either within the numor files in the /rawdata/ subdirectory corresponding to the performed experiment, within auxilliary files such as logfiles or parameter surveys stored respectively in the /logfiles/ or /histo/ subdirectories, or within the /processed/ subdirectory when created or uploaded by the authorized members of the proposal. Proposal-related metadata (sample description, sample environment, etc) will be curated in the ILL proposal database. Electronic metadata other than that produced by beamtime instrument computers (e.g. from computers in auxillary ILL laboratories), as well as non-electronic metadata (e.g. instrument notebooks), are in general not fully curated by the ILL.

3.2.3 Processed data (i.e. resulting from some level of data treatment) will be curated as files of various formats (depending on the data treatment software used) in the /processed/ subdirectory corresponding to the performed experiment.

3.2.4 Several days after the end of the experiment(s) corresponding to a given beamtime proposal, a DOI (Digital Object Identifier) is generated that incorporates the proposal number and is linked to all the data for that proposal (raw data, metadata, processed data) via the ILL Data Portal (data.ill.eu).

3.2.5 Accepted beamtime proposals having scheduled but not yet performed experiments (and therefore no DOI yet) have empty data directories that can nevertheless be accessed via the ILL Data Portal using the beamtime proposal number and/or other searchable keywords.

3.2.6 All data will be migrated or copied to archival facilities for long-term curation.

3.3 Access to data curated at the ILL

3.3.1 Access to ILL data is restricted to ILL staff and registered ILL users, namely those individuals having an ILL User Club account.

3.3.2 The ILL provides each registered user with a personal computer account that allows him/her to use ILL computing facilities, along with a limited amount of personal disk space on ILL disk servers (e.g. /home/notill/x/xyz), and thereby to access data from his/her experiments using ILL workstations via access-control lists (ACLs). In particular, the MyData subdirectory in the user's home directory is symbolically linked to the data file directories of all his/her accepted beamtime proposals. The registered ILL user is responsible for the secure use of his/her ILL computer account.

3.3.3 Authorized access to data (raw, meta- and processed) from outside the ILL's firewall is enabled via the ILL Data Portal (data.ill.eu), and also via remote login to certain ILL workstations subject to authorisation by the user's local contact who then assumes responsibility for the remote access. Once a registered user is logged into an ILL workstation (either from within the ILL's firewall or via remote login) and using ILL disk servers, access to data is controlled by ACLs.

3.3.4 Access to all data pertaining to an ILL experiment, whether via the ILL Data Portal or ACLs, is restricted to the authorized members of the corresponding ILL beamtime proposal for a period of 3 years after the end of all the associated experiment(s). If data can be stored at the ILL for only less than 3 years (e.g. due to capacity limitations), then access is exclusive to the authorized members up to the end of the storage period.

3.3.5 From 3 years to 5 years after the end of the experiment(s) corresponding to an ILL proposal (or to a CRG proposal making use of ILL curation tools for the data), non-authorized members of the proposal who are however registered in the ILL User Club may request access to the data, and in such case, the ILL management in coordination with the PI will come to an agreement about granting access (generally in the form of download access only). Open access is thereby guaranteed because any serious researcher can register in the ILL User Club, and this procedure then provides the mechanism via the ILL Data Portal for informing the PI of downloads and facilitating collaborations and proper use of the data (see below).

3.3.6 Any PI who wishes that access to his data remain restricted for more than 5 years must make a special case to the ILL management, otherwise the data will be released and become open data.

3.3.7 Data can always be released earlier by proposal managers via the ILL Data Portal (data.ill.eu). A limited subset of metadata (sample composition, proposal abstract) will be released when the user-submitted Experimental Report for the proposal has been released (according to a separate procedure lasting 6 to 12 months).

3.3.8 Once data for a given proposal have been released and become open data (an irreversible event), they can be downloaded via the ILL Data Portal only by registered ILL users. Since ACLs are not changed when data are released, all ILL-curated metadata and processed data remain modifiable only by authorized members of the corresponding proposal who retain exclusive write-access either via ACLs or the ILL Data Portal. *Releasing ILL data to become open data therefore grants download access only.* Downloading of open data from the ILL Data Portal will be logged and that information made available to the PI, including the identity of the downloader.

3.3.9 All ILL open data are released under the terms of the Creative Commons license CC-BY (http://creativecommons.org/licenses/) which must be adhered to by the downloader.

3.3.10 Non-electronic metadata (e.g. instrument notebooks, other notes made by the PI), as well as electronic metadata other than that produced by beamtime instrument computers (e.g. from computers in auxillary ILL laboratories), all of which are in general not fully curated by the ILL, become open data “indirectly” in that the downloader is encouraged to contact the PI and to suggest collaboration. The PI then has the option to share such metadata that he possesses with the downloader, but the PI is not obliged to do so.

3.3.11 Whether during or after the embargo period, authorized members of a proposal always have full access to the data of each of their beamtime proposals. Concretely this means: read-only access to the /rawdata/, /logfiles/ and /histo/ subdirectories corresponding to the proposal, and read+write access to the /processed/ subdirectory including any of its subdirectories as created by the authorized members. Note that authorized members cannot create additional files nor subdirectories at the same level as /processed/, so that /processed/ remains their only write-accessible subdirectory for the proposal.

3.3.12 ILL-curated raw data will be read-only (i.e. non-modifiable) for the duration of its lifetime.

3.3.13 ILL-curated metadata will generally be read+write-able (only by authorized members of the proposal), but read-only for that within raw data files, beamtime proposal files, and certain logfiles.

3.3.14 ILL-curated processed data are generated from raw data and/or metadata and are therefore necessarily read+write-able, but only by authorized members of the proposal.

3.3.15 Only authorized members are permitted, both during and after the embargo period, to upload additional metadata or processed data to the /processed/ subdirectory corresponding to a proposal's experiment, either via the ILL Data Portal (data.ill.eu) or by using ILL workstations. Authorized members of a proposal whose User Club accounts have been authorized for remote access (see 3.3.3 above) can additionally make use of tools like sftp (secure file transfer protocol) for uploading files to /processed/ from outside the ILL's firewall. Metadata or processed data should not be removed or changed in any way that could degrade the usability of the data by an interested party.

3.3.16 It is the responsibility of the PI to ensure that the proposal number is correctly entered into the instrument control computer at the start of the experiment, so that data acquired during the experiment will be saved under the proper identifier. (If this is not done, the authorized members might not be able to access the data or other users may inadvertently be given access rights to the data, at least until the ILL's central computing services can correct the mistake.)

3.3.17 Appropriate ILL staff (e.g. instrument scientists, support staff) have access to ILL-curated data for logistical reasons. The ILL will undertake to preserve the confidentiality of such data.

3.3.18 The ILL Data Portal enables the linking of ILL data to the corresponding beamtime proposal, as well as to any resulting publications, via the proposal number (and/or the corresponding unique DOI) associated to the proposal. Access to the electronic files of submitted beamtime proposals will however never become open, since the proposal's scientific description is probably not up-to-date, hence possibly misleading, and therefore less useful than the content of publications resulting from the proposal.

3.3.19 The PI and other proposal managers have the right to transfer or grant parts or all of his/her rights to another authorized member of the proposal, and to add new authorized members from the pool of registered ILL users (i.e. those having an ILL User Club account). (The corresponding ACLs are updated within a few minutes.)

3.3.20 The PI has the right to create and distribute copies of data from his/her experiments, even during the embargo period.

4. Policy towards results obtained from ILL data

4.1 Ownership of results

4.1.1 Ownership (in the sense of intellectual property) of all results derived from the analysis of ILL data is determined by the contractual obligations of the person(s) performing the data analysis and/or data interpretation. Note that results can be produced either during the embargo period by authorized members of the proposal (and their collaborators), or afterwards by other registered ILL users and their collaborators who have made use of ILL open data.

4.1.2 Note that any results that are stored in the /processed/ subdirectory of a performed experiment, or in any of its subdirectories as created by the authorized members, are treated as processed data and thus destined to become open data after the embargo period.

4.2 Curation of results by the ILL

4.2.1 The curation of results by the ILL is limited to those files stored in the /processed/ subdirectory (including any subdirectories of that subdirectory) corresponding to a performed experiment. Even though authorized members can create subdirectories of /processed/ for storing results, they do not have write-access for creating additional subdirectories at the same level as /processed/, nor for modifying the content of the read-only subdirectories /rawdata/, /histo/ and /logfiles/.

4.2.2 The limited amount of personal diskspace provided by the ILL to a registered user (see section 3.3 above) is not accessible via the ILL Data Portal and therefore should not be used to store results pertaining to a performed experiment.

4.3 Access to results curated at the ILL

4.3.1 The access to results curated at the ILL is equivalent to the access to processed data curated at the ILL, as both are stored within the /processed/ subdirectory of the performed experiment, and both will be released and enjoy open access after the end of the embargo period.

4.3.2 The use of ILL workstations from within the ILL firewall, as well as tools like sftp (see 3.3.15 above) and soon the ILL Data Portal (data.ill.eu) from outside the firewall, will allow authorized members to continue to upload results to the /processed/ subdirectory of a performed experiment even after the data become open data at the end of the embargo period.

5. Guidelines for management of metadata and results

5.1 The authorized members of a proposal are encouraged to ensure that the metadata they supply are as complete as possible, including (for example) concise but sufficiently descriptive sample designations, since such practices will enhance the possibilities for themselves, as well as for other interested parties, to search for, retrieve and interpret the curated data in the future.

5.2 The ILL undertakes to provide means to the authorized members via e.g. the ILL Data Portal for the uploading of such metadata items that are not automatically captured by an instrument, in order to facilitate recording the fullest possible description of the raw data corresonding to a proposal. In general, such supplementary and/or complementary metadata will be uploaded to the /processed/ subdirectory where they will be curated and then become open data after the embargo period.

5.3 Since releasing ILL data as open data offers download access only, any treatment or processing of raw data and metadata that involves write-access to the /processed/ subdirectory remains controlled by the managers of the corresponding proposal, both during and after the embargo period.

5.4 **A researcher who aims to produce results from ILL open data should, when possible, contact the PI to inform them and suggest a collaboration if appropriate.** In any case, he/she must acknowledge the ILL beamtime proposal as the source of the data and cite its unique identifier (DOI) and any publications of the PI that are linked to those data. A reminder of these obligations is posted when open data are downloaded via the ILL Data Portal (data.ill.eu).

5.5 The downloading of open data via the ILL Data Portal also displays a disclaimer stating that the ILL cannot be held responsible for the misuse, misinterpretation, or misrepresentation of open data that were released under the ILL's data policy.

5.6 Researchers who produce results from ILL open data are encouraged to link those results to the corresponding ILL beamtime proposal by using the ILL Data Portal. Authorized members are further encouraged to upload their results from ILL open data to the /processed/ subdirectory of the corresponding proposal so that such results enjoy open access as well.

6. Publication information

6.1 Anyone publishing results based on data from an ILL beamtime proposal, either before or after the data have become open data, **must cite the DOI corresponding to the ILL beamtime proposal in the references section of the publication** (and in related publications if appropriate).

6.2 References (i.e. information for making a citation) for publications related to experiments carried out at the ILL must be submitted (e.g. via email) to the ILL's publications database (maintained by on-site library staff: library@ill.eu) within 3 months of the publication date, or during any new application for beamtime, whichever is the earlier.

6.3 It is encouraged that preprints, postprints and reprints of publications related to ILL experiments be promptly submitted to the library staff.