T H R I L L

TECHNOLOGY FOR HIGH-REPETITION-RATE INTENSE LASER LABORATORIES

Deliverable Data	
Deliverable number	D2.2
Deliverable name	Data Management Plan
Work Package	WP2
Lead WP/deliverable beneficiary	Laserlab-Europe AISBL (LLE-AISBL)
Type and dissemination level	Report, public
Deliverable status	
Submitting author	M. Fischer
Verified (WP leader)	Management board
Approved (Coordinator)	
Due date of deliverable	30.06.2023

Table of Contents

Funded by the European Union

This project has received funding by the European Union's HORIZON-INFRA-2022-TECH-01 call under grant agreement number 101095207

List of tables

About THRILL

The THRILL project deals with providing new schemes and devices for pushing forward the limits of research infrastructures (RI) of European relevance and ESFRI landmarks. To do so, the project partners have identified several technical bottlenecks in high-energy high-repetition-rate laser technology that prevent it from reaching the technical readiness level required to technically specify and build the needed devices, and guaranteeing sustainable and reliable operation of such laser beamlines at the partnering RIs. Advancing the technical readiness of these topics is strategically aligned with the long-term plans and evolution of the ESFRI landmarks FAIR[1], ELI (-BL) and Eu-XFEL, and RI APOLLON, bringing them to the next level of development and strengthening their leading position.

The project is focused and deliberately restricted to three enabling technologies, which require the most urgent efforts and timely attention by the community: high-energy high-repetition-rate amplification, high-energy beam transport and optical coating resilience for large optics. To reach our goals, the major activity within THRILL will be organized around producing several prototypes demonstrating a high level of technical readiness. Our proposal is addressing not yet explored technical bottlenecks – such as transport over long distances of large-aperture laser beams via relay imaging using all-reflective optics – and aims at proposing concrete steps to increase the performances and effectiveness of the industrial community through the co-development of advanced technologies up to prototyping in operational environments.

The project is not only pushing technology, it is also offering an outstanding opportunity to train a qualified work force for RIs and industry. With this in mind, the structure of THRILL promotes synergetic work, fast transfer to industry and integrated research activities at the European level. Access to the RIs will be granted as in-kind contribution.

[1] Here FAIR stands for the “Facility for Antiproton and Ion Research”, a new research infrastructure being built in Darmstadt (Germany) - not to be confused with the “FAIR” data approach later in this document.

Executive summary

The purpose of Deliverable 2.2 “Data Management Plan” (DMP) is to determine how the project’s research data will be handled both during and after the project. This includes the collection, process and generation of data within the THRILL project. The DMP provides key actions and strategies to ensure that the research data is in line with the FAIR principles: Findable, Accessible, Interoperable and Reusable. It describes the types of data the project will collect, how the data will be stored, and how it will be made available for validation, exploitation and re-use by partner organisations.

Abbreviations

Abbreviation	Definition
DMP	Data Management Plan
FAIR (data approach)	Findable, Accessible, Interoperable and Re-usable data approach
FAIR (facility)	Facility for Antiproton and Ion Research in Europe
GA	Grant Agreement
ML	Machine Learning
RI	Research Infrastructure
THRILL	Technology for High-Repetition Rate Intense Laser Laboratories
WP	Work Package

1 Introduction and objectives

THRILL is a technical project that aims at advancing knowledge on high-energy lasers on topics of high relevance for the scientific community. The project structure builds on the complementary expertise of the partners, and therefore the experiments performed during the project require cross-techniques and multi-facility exploitation, together with a good data management plan for both internal and external communication.

As a consequence, the THRILL consortium considers open science as a crucial facilitator for accelerating research impact by ensuring transparency of research, research integrity, and the transfer of knowledge to industry and to society in general. The consortium will ensure that the open access requirements are broadly known and will strongly encourage compliance by all partners.

The aim of this Data Management Plan (DMP) is to specify how the project's research data will be handled both during and after the project. This includes the collection, process and generation of data within the THRILL project. The DMP provides key actions and strategies to ensure that the research data is in line with the FAIR principles: Findable, Accessible, Interoperable and Reusable. It describes the types of data the project will collect, how the data will be stored, and how it will be made available for validation, exploitation and re-use by partner organisations. It further considers the different requirements of the different infrastructures and user groups. The DMP is developed based on the DMP template provided by the European Commission for Horizon Europe projects.[2]

The THRILL DMP is a living document that will be updated regularly.

[2] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan-template_he_en.docx (8 May 2023)

2 Data Management in THRILL

2.1 Data summary

Although in the abstract sense, every research effort can be summarized by the process of acquiring and analysing of information, i.e. data, the project partners find it important to point out that there is a considerable range in the practical implementation of this concept. While in some fields of research, it is common practise to analyse and interpret well-structured data generated by standardised methods and devices, this might be more the exception in fields dedicated to the development of such data sources. As has been outlined, THRILL operates at the forefront of the development and improvement of laser technology, which will naturally involve a considerable amount of concept and prototype testing. In many cases, this will lead to situations where the raw data, e.g. generated by a measurement device which is either a commercial device or a setup which has been developed on the spot for this purpose, is far less important than its interpretation and the resulting decisions on the further development. And as a consequence, the interpretation of the term “data” in the sense of “information and documentation of the process of acquiring it” seems to be more appropriate than e.g. “as much raw data as technically possible”. A monitoring process has been established for the careful assessment of the respective situations, with the aim of finding a good compromise between depth of documentation including the necessary effort and the value to the concerned research community.

Coming back to a more structural description, THRILL will generate two major categories of data: documents related to the management and communication of the project on the one hand, and on the other hand all scientific data from Work Packages 3, 4, 5 and 6 governed by the FAIR principles.

A summary of the envisioned types of data generated in the project can be found in Table 2. At this early stage in the project, there is only a rough estimation of the size of the data given. The table will be updated along the course of the project.

Table 1 – Summary of data generated in the THRILL project

Data type	Contents	Format	Data size	Archive
Text files, PDF documents	Minutes, deliverables, milestones, reports, presentations, email distribution lists, lists of participants, templates, training and workshop material, technical notes and reports	txt, pdf	100s of Megabytes	GSI cloud THRILL website
Images	Photos, logos, graphics	jpg, png, gif, svg, pdf	100s of Megabytes	GSI cloud THRILL website
Experimental data	Numerical tables, images, text, scans, methodological information, technical drawings, optics calculations and measurement reports	csv, txt, markdown/markup, open image formats, FITS cubes	Up to 500 GB per task	Local storage / servers
Publications	Scientific publications, journal articles	pdf	100s of Megabytes	THRILL website Journals Zenodo
Source code	Source code developed in the development / deployment / implementation process		10s of Megabytes	Tbd., e.g. GitLab, GitHub

In addition to the newly generated data, in the context of experiments realized for THRILL, data, source code and technical drawings of preliminary experiments and other projects will be used. Documenting the work done for the project, for publications and for the project deliverables will generate data in the form described in table 2. Some data will come from test runs done on the different laser facilities of the laboratories (e.g. ATONLULI2000, Apollon) and from laser development done in the context of the THRILL project. The experimental data, source code and documents will be useful for a wide range of scientists who are developing a beamline or experiment for such a laser or similar laser technology.

2.2 FAIR data approach

The experiments performed during the THRILL project are of high complexity and encourage and require cross-techniques and multi-facility exploitation. The resulting data need to respect the FAIR data approach as broadly as reasonably possible according to the EU's open science policy. The FAIR principles state that data must be "findable", "accessible", "interoperable" and "re-usable".

The following sections of the DMP lay out the methodology followed in the framework of THRILL with respect to the FAIR data approach.

Making data findable

All published data will be assigned unique permanent identifiers (DOI) through publication in public domain or institutional repositories. The creation of metadata depends on the used data storage system but a common set of metadata for the project has been agreed upon by the partners. Institutional repositories provide usually the environment to describe data with rich metadata. This includes for instance keyword indexing and the presentation of metadata in a way that can be harvested and indexed. For metadata not included in the repository structure, accompanying documents (readme-files) describing the recording, composition and previous processing steps of the data sets will complete this. Given the nature of the data produced, the following metadata will be systematically produced and indexed, which should be enough to describe the data:

- Type of document
- Authors and affiliation
- Date of generation
- Relevance in project (work packages and sub-task)
- Keywords

For a lot of tasks within THRILL, the number of data sets will be project-specific and therefore not map out a large parameter space which could be integrated in existing data repositories. All data will be mentioned in associated publications, and a table-of-contents text-file, describing the composition of the data sets, will be added to the repository.

For any produced data, the scientific data policies of the beneficiaries apply:

- **GSI and FAIR facility:** https://doi.org/10.15120/GSI-2023-00646
- **HZDR:** https://www.hzdr.de/db/Cms?pOid=57725
- **CNRS:** will follow the same process as for the Apollon laser infrastructure: https://apollonlaserfacility.cnrs.fr/wp-content/uploads/2022/06/APOLLON-data-policy.pdf
- **ELI:** https://zenodo.org/record/6515903
- **EU XFEL:** https://www.xfel.eu/users/policies/index_eng.html

Especially for the instances of data and information not included in the deliverables and publications, we want to establish a low-entry-level reporting system based on form sheets where task leaders can supply informal, few-paragraph reports and links to associated data. These reports will be published on the project website in areas dedicated to the work packages, and the respective work package leaders will encourage the other project members to use this instrument to report their findings. This approach deliberately differs from the publication in peer-reviewed journals and is explicitly meant also for preliminary findings which can be amended or corrected at a later time.

Making data accessible

Repository

The resulting data of the investigations performed in THRILL need to respect the FAIR data approach as broadly as possible according to the EU's open science policy, while at the same time taking into account intellectual property rights issues of the participating industrial partners in order to ensure potential commercial exploitation of the results.

For publications resulting from the THRILL projects, data will be made available in trusted repositories. Well-established institutional data repositories at the participating facilities will ensure the long-term preservation and curation of the scientific data generated by the project.

Within the THRILL consortium, the following repositories are used:

- **GSI and FAIR facility:** http://repository.gsi.de/
- **EU XFEL and HZDR:** RODARE, the Rossendorf Data Repository, https://rodare.hzdr.de/
- **CNRS:** HAL, https://hal.archives-ouvertes.fr
- **ELI:** Repository pending, in preparation

These repositories have been maintained for many years and the consortium members have experience in using them (DOI assignment, metadata). In particular, these are well adapted for the type of data produced by the project. In addition, a dedicated THRILL community will be set up on zenodo.org to ensure the findability and accessibility of the data and publications resulting from experiments implemented in THRILL.

THRILL will encourage open access publication with a preference for the gold route to open access, but acknowledges that there may be work published according to the green one. For publications of scientific results arising from the project itself, journals with gold open access and with high impact factors like HPLSE or Optics Express, or the open access publishing platform "Open Research Europe" will be considered with high priority. In the case of exceptional results meriting inclusion in very high impact publications without the option of open access, then open access (green open access) will be ensured by archival of the peer-reviewed preprints in an online repository, either in partner institutional archives (such as HAL in France) or in disciplinary archives as Zenodo (or arXiv). The same repositories will be used to archive gold open access papers. Authors will be asked to provide proper justification for choosing publication routes in hybrid open access journals.

As stated in the HE Annotated Grant Agreement, beneficiaries will ensure open access to peer-reviewed scientific publications relating to their results. Immediate open access will be ensured at the same time as the first publication, if needed through a trusted repository.

As a next step, guidelines on the Horizon Europe open science requirements will be set up and

shared with the researchers involved to guarantee the alignment of the publications with these requirements directly from the beginning of the project.

Data

As for now, all data will be openly available. The type of data that will be made available by THRILL is mostly of documentation nature, made of reports, technical notes, software and drawings.

For software, e.g. the modules for the “cacao” project that will enable high-speed real-time evaluation of Shack-Hartmann sensors on GPUs, the development will be made under the GNU license and the software development will use the GIT environment (e.g. https://github.com/cacao-org).

However, the information gathered in some cases, e.g. during the testing of commercial developments, will only be shared out of the consortium after the involved company has agreed. In addition, there are some practical concerns as the raw data of Task 5.2a may be of 100s of GBs, which greatly elevates the necessary effort of making them publicly available. The project partners prefer to only make data publicly available after intermediary analysis (with according descriptions), which will be of higher scientific interest.

Given that THRILL works in an environment where access to facilities for gathering data is scarce and subject to lengthy application and planning processes, complete sets of data may take many years between data taking, analysis and waiting for the next data taking cycle. Especially for data which is also to be exploited for scientific publication, an embargo of up to 3 years can be implemented to make sure that researchers have time to do the additional experiments and analysis needed to publish, knowing that most of the experiments are done on facilities where time is allocated once a year or on even longer intervals. This arrangement is vital to ensure that the involved researchers - often graduates and post-graduate students - can exploit the data they have created in their work in a competitive research field in an adequate way.

Given that the data will be made available through institutional repositories, access and the access protocol will be that of the repositories, e.g. https. If there are restrictions on the use of the data, the data will be accessible with a password given by the laboratory during the embargo time. For CNRS, some data will be accessible with mandatory registration.

The beneficiaries agreed that for now, there is no need for a data access committee, and that the question of permanent access to restricted data and identification will be dealt with if the need arises during the course of the project.

Metadata

The metadata will be made publicly available through the hosting repositories at the same place as the data themselves and the metadata's lifetime will be as long as the one of the actual data as they will be hosted on the same repository.

If any software is necessary to read or access the data, links will be provided to the public repositories, accompanied with the last version number which was known to work. The beneficiaries do not plan to include specific software, instead they will attempt to use established open data formats which do not rely on a specific implementation.

Making data interoperable

The project outputs will use open data formats wherever possible (pdf, csv, txt, markdown/markup, open image formats), depending on what is currently used at the respective RIs. All data will be specified as necessary for interoperability in accompanying publications and metadata documents.

Increase data re-use

Data, text, and images will be – wherever possible – published using Creative Common licenses (by default CC-BY 4.0 and its derivatives as needed). Software will be made available using very permissive software licenses (e.g. Apache, MIT, BSD) unless the extension of existing software or industry transfer requires a more restrictive licence.

Data will be made freely available in the public domain to permit the widest re-use possible. The data will be provided in a way that it can be used by third parties.

Readme-files will be included. If it is decided that raw data are not made publicly available but processed/reduced data, raw data samples will be included with our processing outcome to facilitate verification.

The provenance of the data will be thoroughly documented in the form of readme files and in the accompanying publications.

Data quality assurance processes will be in place. However, due to the differences in the types of data acquisition, the means of verifying the quality vary strongly and have to be specified at a later date.

3 Allocation of resources

Project outputs will be directly integrated into the data curation/preservation procedures of the partner RIs, which include financial and technical plans for preserving published/generated data as part of the facilities' operational costs. Curation of meta-data will be handled by the scientific libraries associated to the project partners as they also handle (data) publication processes and provide permanent identifiers (DOIs).

Where applicable, the cost related to open access publications will be covered by the project, where a corresponding budget has been foreseen. In addition, the partners have access to institutional funding for open access publication, would that turn out to be insufficient.

Resources for long term preservation are discussed within THRILL. The cost of curating and storing the given amount of data will be studied.

4 Data security

All data generated by THRILL do not require specific data protection measures, as it would be the case if the data would involve personal information and fall under the European data protection legislation.

At the partner institutions, data are systematically saved on internal and protected servers with a daily back-up of hard drives. Some partners are using open-source cloud software and local RAID storage technologies. According to local regulations, primary data used for publications sometimes has to be securely stored for a defined period of time, e.g. ten years, in a durable form in the institution of origin.

Central data repositories for open access mostly are managed by institutional departments, e.g. libraries, which ensure data security - including protection against malevolent hacking – as well as reliability and sustainability. HAL (CNRS) is for instance going to be certified by CoreTrustSeal to demonstrate it is a trustworthy data infrastructure.

5 Ethics

For now, the beneficiaries do not expect any ethical or legal issues impacting data sharing and do not deal with personal data. If according questionnaires will be created, informed consent for data sharing and long term preservation will be included.

6 Conclusion

This document describes the collection, process and generation of data within the THRILL project. The key actions and strategies of the DMP were presented, to ensure that the research data is in line with the FAIR principles: Findable, Accessible, Interoperable and Reusable. It is described which types of data the project will collect, how the data will be stored, and how it will be made available for validation, exploitation and re-use by partner organisations.

Over the course of the THRILL project, the DMP will be updated regularly whenever the need arises as we see it as an active document.