Project Number: 766719

Start Date of Project: 2017/11/01

Duration: 36 months

Deliverable 1.3 – V1.0

Title: Data Management Plan

Dissemination level	Public
Submission Date	2018/04/23
Work Package	WP1-Management
Task	1.1
Type	Report
Version	1.0
Author	Luca Persichetti
Reviewed by	Giovanni Capellini
Approved by (CO)	Monica De Seta

Table of Contents

List of abbreviations

FLASH	Far-Infrared Lasers Assembled usung Silicon Heterostructures
DMP	Data Management Plan
M	Time in months after the starting date of the project
THz	TeraHertz
Si	Silicon
CMOS	Complementary metal-oxide semiconductor
Ge	Germanium
mW	Milliwatt
OpenAIRE	Open Access Infrastructure for Research in Europe

List of figures

List of tables

1. Executive Summary

This document is a deliverable of the FLASH Project which is funded by the European Union's H2020 Programme under Grant Agreement No. 766719. This first version of the Data Management Plan (DMP) describes the main elements of the data management policy that will be used by the members of the Consortium with regard to the data generated throughout the duration of the project.

The DMP is released in compliance with the H2020 FAIR (making data Findable, Accessible and Interoperable) principle 103 and will be updated at months M18 and M36.

The data generated in FLASH will be mainly experimental and characterization data, design data, computational modeling data, publications and project documents and reports. FLASH will ensure open access to all the data necessary to validate the research results of the project, including publications. Research data linked to exploitable results will not be immediately available in the open domain in case this may compromise its commercialization prospects. If needed, embargo periods may be defined and specified in the future versions of this document.

2. Scope of the deliverable

The purpose of this document is to provide the plan for managing the data generated and collected during the project. The DMP describes the data management life cycle for all datasets to be collected, processed and/or generated by FLASH. It covers:

- the handling of research data during and after the project;
- what data will be collected, processed or generated;
- what methodology and standards will be applied;
- whether data will be shared/made open and how;
- how data will be curated and preserved Following the EU's guidelines regarding the DMP.

The DMP is a live document, updated during the project as illustrated in Figure 1. We assume three incremental releases of the DMP: After the initial release at M6, at months M18 and M36 (end of the project), respectively.
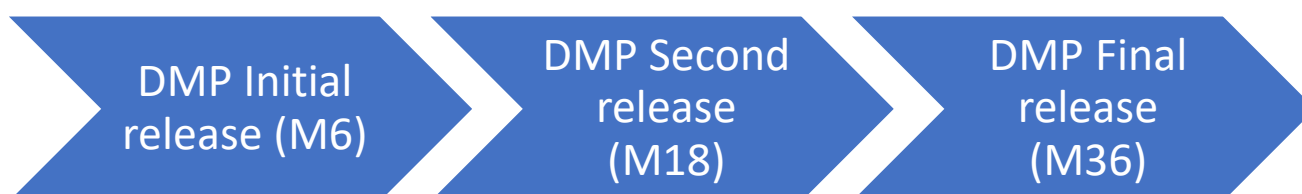

Figure 1: DMP Life Cycle.

Any new version of the DMP will include all the information of the previous release, together with the necessary updates/corrections. After the release data of the new DMP, the information contained in the previous versions will be considered obsolete.

3. FLASH DMP

FLASH is a project that aims to develop a room-temperature THz laser integrated on Si using CMOS technology-compatible processes and materials. The laser, of quantum-cascade type, will be assembled using newly developed conduction-band germanium-rich heterostructures. It will leverage on the non-polar nature of Si and Ge crystal lattices to potentially enable room-temperature operation and will emit > 1 mW power in the 1-10 THz range. The members of the Consortium working on the project are: 1. Laboratory of Mesoscopic Physics and Nanostructures-Department of Science, Roma Tre University (Italy); 2. University of Glasgow-UGLA; 3. Innovations for high performance microelectronics/Leibniz-Institute-IHP; 4. Quantum Optoelectronic Group- Department of Physics ETH Zurich-ETH; 5. Nextnano GmbH-NXT.

FLASH vision of a DMP is inspired by the FAIR data Principles [1]. Therefore, FLASH will consider the following approaches as far as applicable for providing the open access to research data:

- published data available on the Web (whatever format) under an open license;
- use of non-proprietary formats (e.g. CSV or TSV)
- use of Metadata and Fact sheet to denote data
- link data to other data to provide context

The policy for open-access to the research data in FLASH will follow the basic principle: "as open as possible, as closed as necessary" which can be translated into two core principles:

1. The generated research data should generally be made as widely accessible as possible in a timely and responsible manner;
2. The research process should not be impaired or damaged by the inappropriate release of such data.

The FLASH consortium shall implement procedures that are in line with national legislation of each consortium partner and in line with the European Union standards. This DMP will apply to all data under FLASH consortium control. If we shall strive to make data open, we cannot overrule limitations that partner institutions put on data that they contribute (see e.g. the grant agreement).

The Governing Board (GB) will assess under strict criteria the nature of the data and will give advice in order to establish their categorization as **open**, **embargo,** or **restricted,** taking into account that the main aim to disseminate the results of the research should be balanced with the necessity to protect the interests of the partners involved in the project.

In that sense, data sets containing key information that could be patented for commercial or industrial exploitation will be excluded for public distribution and data sets containing key information that could be used by the research team for publications will not be shared until the embargo period applied by the publisher is over.

In particular, the open access to the research data can be denied in cases of:

• the results are commercially or industrially exploited,

• incompatibility with confidentiality and security issues

• protection of personal data – privacy,

• the disclosure is likely to jeopardise the achievement of the main aim of the action,

• other legitimate reason.

In the following, we define as **data set** either an individual file (such the pdf file containing a deliverable or a report) or an ensemble of files which are logically connected, like different measurements of the same sample or sharing some physical observable. In the latter case, the files will be bundled in a single zip file which will be labelled with a single metadata etiquette.

All the data sets, regardless of their categorization, will be stored in the consortium private-access data repository accessible through the SFTP/SSH protocol. In addition, those categorized as open or embargo will be publicly shared (in the case of embargo, after the embargo period is over) through the public section of the project website (www.flash-project.eu) and ZENODO (https://zenodo.org/), an open access repository for all fields of science that allows uploading any kind of data file formats, which is recommended by the Open Access Infrastructure for Research in Europe (OpenAIRE).

For all the data sets a Fact sheet will be filled by the authors in order to summarize the characteristics of each data set to give a quick understanding of the content of the data to anyone that reads it.

In order to make those data sets that are publicly shared as discoverable and accessible as possible the following aspects will be considered:

- The **Metadata** will have a key role in the improvement of the discoverability of what we upload to be open to the general public. Taking into account the huge amounts of information that can be found on the Internet, it is necessary to use a standard set of encoded labels in the websites where we store our data in order to make it easier for the search engines used by the browsers (such as Google, Bing, Yahoo…) to find it. There will be three types of metadata for each public data set: common metadata (related to the EU, H2020, name and number of the project), specific metadata (3 Keywords chosen by the authors of the data) and fact sheet metadata (related to the nature, origin, description, authors, potential interested groups, etc.). Considering the amount of data that will be foreseeably produced in the project, we shall gather similar data in compressed files, sharing common Fact sheet and Metadata (e.g. spectroscopic data of devices for different excitations)

- The **Standardization** of the names and formats of the files stored and uploaded will also improve the accessibility to the information, as we will describe in the following.

The FLASH consortium will take, within the boundaries above stated, the appropriate measures so that the research data generated in the project is easily discoverable, accessible, assessable and intelligible, useable beyond the original purpose for which it was collected and interoperable to specific quality standards.

4. Metadata and standardization in FLASH

In the context of this document, metadata is organized information labelling a data set and encoded in the code of the websites in order to facilitate discovery and reuse of the information by third parties.

Three types of metadata will be defined for each data set:

1. **Fact sheet information**: As stated in the Section "5. Fact sheet information" of this document, for each data set the authors will have to fill a Fact sheet that allows anyone to quickly identify the content of the data set.

2. **Common metadata**: According to the "Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020" regarding the research data generated, the beneficiaries of the grants should follow Article 29.3 of the Grant Agreement which states that the bibliographic metadata must be in a standard format and must include all of the following terms:

a. European Union (EU);
b. Horizon 2020;
c. Name of the project: Far-Infrared Lasers Assembled using Silicon Heterostructures;
d. Acronym: FLASH;
e. Grant number: 766719.

3. **Specific metadata**: The authors will have the option to choose up to 3 Keywords that they consider relevant for the data set and can be of frequent use if someone is searching for the kind of data contained on the data set. Once the Fact sheet is fulfilled it will be sent with the data set to the Website managers. The Website managers will use the information indicated by the authors to complete the metadata of the data sets that are going to go public. The metadata will not be use in those data sets that has been categorized in the Fact sheet as "Restricted".

In order to make the information accessible for internal and external users and according to the good practices for "Open data" free file formats such as PDF, OpenOffice, PNG (portable network graphics) and SVG (scalable vector graphics) will be prioritized when uploading information.

Regarding the names of the files, short descriptive and consistent file names will be key to make it easier to locate the information needed now and in the future. The rules to name data set files are reported in Table 1:

Convention	[time_stamp]_FLASH_[data type]_[data postfix]_[version].[file format]				
Item	Time Stamp	Data Type	Data type postfix	version	File format
Optional	X		X	X	
Definition	YYYT_MM_DD	Design Simulation Measurement Publication Document Report Deliverable Presentation	arbitrary	v#.#	According to software: pdf jpg zip xlsx docx pptx ...
			Specific Equipment used for acquiring, e.g. AFM, SEM, etc.		
			Project Meeting		
			Type of model, Simulation, etc.		
Examples	2018_03_31	DelirevableReport	D1.1	v1.1	As above
Complete	*2018_03_31_FLASH_DelirevableReport_D1.1_v1.1.pdf*				

Table 2: FLASH data file naming convention.

5. Fact Sheet information

For each data set the researcher will fill the Fact Sheet shown in Table 2. The fields specified in that Fact Sheet should be filled according to the following rules and recommendations.

5.1 Data set description

5.1.1 Reference

Each data set will have a reference that will be generated by the combination of the name of the project, the Work Package and Task in which it is generated and a consecutive number (15 characters maximum, for example: FLASH_T1.0_01).

5.1.2 Description

An intelligible description of the data collected, understandable for people that do not directly work in the project, and independent from other data set descriptions, so it can be understood without having to go through every data set. (60 characters maximum).

5.1.3 Authors

The name of the Authors and the Entity will have to be completed.

5.1.4 Origin

The researchers will have to select the origin or origins of the data between the next options:

- Laboratory experimental data;
- Computer simulation;
- Review;
- Design, drawings;
. Papers;
- Other, to be specify.

5.1.5 Nature

The researchers will have to select the nature of the data between the next options:

- Documents (text, Word processors), spreadsheets;
- Laboratory notebooks, ;
- Type of model, simulation (i.e. Nextnano simulation of wavefunctions);
- Data type based on the equipment used (e.g. XRD, FTIR data)

5.1.6 Sharing Status

The researchers will have to select the sharing status between the next options:

- Open: Open for public disposal.

- Embargo: When a data set is published in a journal, it will become public following the embargo policies of the publisher.

- Restricted: Only for project internal use.

5.1.7 Required software for opening the file

The researchers will have to specify the software required or suggested for opening the file.

5.1.8 Whether it underpins a scientific publication

The researchers will have to answer "Yes" or "No", and in case the answer is "Yes" they will have to give the reference and date to the mentioned publication in the following format: *"NAME OF THE PUBLICATION.Year of publication. DOI".*

Description	Reference	
	Description	
	Authors	
	Origin	
	Nature	
	Sharing status	
	Required Software	
	Whether underpins publication	
Metadata	Common metadata	European Union (EU)
		Horizon 2020
		Far-Infrared Lasers Assembled using Silicon Heterostructures
		FLASH
		Grant number: 766719
	Specific metadata	

Table 2: *Example of Data Fact Sheet*.

6. Type of data in FLASH

This section gives and overview about the research data which are generated, collected, processed and stored in FLASH. This includes the data description for different types and formats, purpose with respect to project objectives and tasks, for data re-use and how, data origin, expected data size, and to whom it might be useful.

Data generated in FLASH will be strictly digital. In general, the data file formats to be used shall meet the following criteria:

• widely used and accepted as best practice within the specific discipline,
• self-documenting, i.e. the digital file itself can include useful metadata,
• independent from specific platforms, hardware or software.

However, different types of data will be generated and handled in FLASH. Considering the technical disciplines related to FLASH project, high-technology equipment and processes are used. Therefore, most of the research data will be in appropriate formats. In the following the main data types are described:

Design data

Design results are schematics and layouts of the active material and of device components. The digital format of the designs is mostly appropriate (e.g. gds files) to the special software (e.g. AutoCAD). Therefore, this software is required for data re-use. Since, these softwares are only available with legal licenses, they cannot be provided by consortium. This makes it impossible to provide an open access to those design data. However, it will be possible to provide pictures and screenshots from the designs which will be collected in design reports or corresponding deliverable reports.

Simulation data

Further data type is generated by collecting results from simulations. These are used to evaluate and estimate the performance and properties of the simulated device. For the simulations, special software is typically used, which need special software licenses for research or even business use. Such software includes for example COMSOL MULTIPHYSICS, MATLAB, Mathematica and the software developed by one of the consortium members (nextnano). The format of simulation results are for a big extent appropriate data set only usable with the simulation software. In some cases simulation results can be exported as ASCII coded text files or to typical database type of formats and spreadsheets with a complete description of the data set (list of fields). These have to be post processed by (simulation data) evaluation software. In most cases are diagrams of the parameters of interest and screenshots the outcomes of simulations. For these reasons it could be not useful to provide the raw simulation data for open access in FLASH. It makes more sense to provide the processed simulation results as diagrams and collected as well as explained in project reports and open-access publications. In case of simulation text files or spreadsheets, the data set could be put in a zip archive and be attached to the report. However, simulation data will only made open after the results have been published.

Measurement data

These data are produced by laboratory experiments and hardware analysis. The data are measurements of specific parameters related to the devices, component, and material properties. Similarly, to simulation data, measurement data are usually appropriate to the used equipment and corresponding software. Sometimes, also typical database types of format or spreadsheets with a complete description of the data set (list of fields) are available.

Therefore, also in case of measurement data, open access of raw data is not worthwhile in FLASH. Detailed measurement data will be made accessible through detailed Supplementary Information sections associated to published articles and, after publication, via measurement reports which also describe in details the measurement environment, e.g. specific performance parameters, test and measurement equipment, experimental setups. If applicable database-based measurement data sets will be attached as zip archive to the reports.

Publications

The most open and visible way to disseminated the data set produced within FLASH will be the publication of research results in scientific articles. Publications are created by one partner for individual results or as

joint publication on joint research efforts and success. Publications are made in scientific journal and on conferences, mostly in the format of pdf files which are commonly usable. In the context of publication data, the open-access approach of Horizon 2020 is embraced by FLASH, following the guidelines presented by the Commission. We shall aim at publishing the project results mainly in fee-based open access scientific journals, following the open-access Gold methodology, and selecting those expected to give the largest visibility of FLASH research activity. For this reason, costs for publication fees have been foreseen in the consortium budget. It is anticipated that FLASH researchers will also primarily target the open-access Green method in the case of conferences and workshop contributions, since the two open-access methods are non-mutually exclusive. In that case the published article or the final peer-reviewed manuscript is archived by the researcher in an online scientific repository (FLASH Website, Arxive, ZENODO repository, etc.) before, after, or alongside its publication. In this case, the authors must ensure open access to the publication within a time frame which is defined by the publisher (embargo times are usually six months to one year).

Project documents ad reports

A second major data set for open access will be project documents and reports, such as deliverable reports. They are generated and collected to summarize the project progress and results as well as discuss different approaches, challenges and deviations with regard to the FLASH objectives. Reports can be related to design, simulation and measurements and contains the processed data. As long as the documents are not assigned to be confidential, contain any confidential data, they are per se public. Normally, documents and reports are in standard pdf format. Public FLASH documents and reports will be made available and accessible on the FLASH webpage after their submission and publication.

7. FAIR data in FLASH

7.1 Making data findable

The research data from this project will be deposited both in:

- A dedicated website for the project: The domain of the website is **https://www.flash-project.eu/.**

- An open access repository: Best practices recommend using an open repository to ensure that the data can be found by anyone. The shared data sets of the FLASH project will be deposited in the ZENOBI repository (https://zenodo.org/). This is one of the free repositories recommended by the Open Access Infrastructure for Research in Europe (OpenAIRE) on their website, and it is an open repository for all fields of science that allows uploading any kind of data file formats.

Both repositories are prepared to share research data in different ways according to how the partners decide the data should be shared:

The dedicated website for the project: Information can be shared in the website at two different levels:

- A private access intranet for internal management of research data.

Each participant of the project will have a username and a password that will be mandatory to enter into the intranet and have access to all the information shared using the SSH/SFTP protocol.

- A public section for the public access to final research data sets. As stated before in this document the data set shall be understood as aggregated data that can be analysed as a whole and has a conclusive and concrete result, and will not include laboratory notebooks, partial data sets, preliminary analyses, drafts of scientific paper. All the information that it is decided to be shared will have no access restriction.

An open access repository: The same Website managers that post the data sets in the public section of the website page from FLASH will simultaneously post it in the open access repository. ZENODO allows to upload files under restricted, open or embargoed access.

7.2 Making data accessible

Generally, all data which after publication can be made accessible. However, as previously described, for most of the data types in FLASH appropriate formats and special restrictions (e.g. software licenses, patents, …) are present which prohibit free-access to the data. For this reason, it has been decided in FLASH that the use of a public repository for raw data is not beneficial at the moment. The data will be stored, however, in the intranet section of the FLASH website. However, a special repository, which links publications to research data, will be evaluated and selected when first data is collected during project progress.

7.3 Interoperable data

In this regard, FLASH will use as wherever possible data formats for open-access knowledge representation, which are formal, accessible, shared and in a broadly applicable language. Qualified references to other data will be included. For example, information on the tools and instruments, which are needed to validate the measurement results, are provided with the data sets. In particular, the format for data sets of equal content, e.g. measurement data, will be a zip archive.

7.4 Re-usable data

Within FLASH. re-usability of data is ensured by the fact that all the data which will be provided for open-access, e.g. documentation, reports papers, will be inherently public (and free of charge). Therefore, no special licenses are established at the moment. However, although the data will public the FLASH project and its consortium members reserve the copyright of the material. For re-usability the data will be stored on the webpage or on a repository system when implemented for at least ten years.

8. Data Security

Data will be stored on the IHP server on which the webpage of the project is located. The IHP repository is managed and support by a team of experts and subject to the institute data security measures and backup. The costs are covered by the IHP budget. Passwords are distributed separately. The original copies of the data will be also stored in the database of the entities that have created them.

9. Ethical Aspects

In the FLASH project, there are no ethical or legal issues present which impairs the data managing. The research in dimension does not create, process and store personal data. Personal data of the FLASH consortium is no subject of the project's data management.

10. References

1. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 2016. **3**: p. 160018.