PHOTONICS PUBLIC PRIVATE PARTNERSHIP

DIMENSION

Directly Modulated Lasers on Silicon

Deliverable Report D1.2

Data Management Plan

Research and Innovation action (RIA)
H2020-ICT-27-2015 Photonics KET

Project Start Date: 1st February 2016
Duration: 48 months
Project reference: 688003

July 31, 2016 – Version 1.0

Project co-funded by the European Commission within
the Horizon 2020 Programme and under
the Photonics Public Private Partnership (PPP) initiative

Dissemination Level: Public

Document information

Title	D1.2 – Data management plan
Work package	WP1 – Management
Responsible	Technische Universität Dresden (TUD)
Due date	Project month 6 (July 2016)
Type	Report
Status	Version 1.0
Security	Public
Authors	R. Henker (TUD)
Project URL	www.dimension-h2020.eu

Confirmation

Any work or result described in this report is either genuinely a result of this project or properly referenced. Any statements and results in the report reflect only the author's view and the European Commission is not responsible for any use that may be made of the contained information.

Version Management

Table 1: List of Revisions.

Version	Description	Author (partner short name)	Released
V1.0	D1.2 draft, final	R. Henker (TUD)	July 31, 2016

Table of contents

List of Figures

List of Tables

Executive Summary

This report describes the handling and open access of research data in the DIMENSION project. Five main types of data is generated, collected and stored: design data, simulation data, measurement data, publication data and project documentation or report data. While the design, simulation and measurement data mostly have appropriate formats which require the use of special software and special hardware, are publications and project documents available in standard pdf format. Furthermore, design simulation and measurement data are restricted to special licenses and NDAs. Therefore, DIMENSION opt-out to enable open access to such data in the current state of the project. Contrary, publications and project documents (as long as not assigned as confidential or inclusions of confidential data are present) are public and open accessible by default. For publications the golden OA way is encouraged and the green OA way is mainly targeted. Since publications and project documents/reports are currently the only data sets considered for OA, no special repository is implemented yet. However, the publications and project documents/reports will be made available via the DIMENSION webpage. In any case the consortium has to agree on the openness of data beforehand.

1 Introduction

1.1 *Open Research Data*

DIMENSION is H2020 project which takes part on the Open Research Data (ORD) pilot following the guidelines given by EC [1], [2]. It is the intention of the EC that each research project ensures that the results, scientific publication and its data behind are open by default (with some reasons for opt-out) and thus provides a broader access to them. This enables several advantages:

- the validation of research results (e.g. for peer reviews) becomes easier,
- scientific breakthroughs becomes more visible,
- research results will be more cited and therefore have a greater impact,
- duplication of research activities will be avoided which improves the quality of results
- research data is preserved,
- EC research funds are better valued and scientific processes become more transparent for society which brings a public benefit, and
- research is better distributed across scientific fields which helps to solve complex (social) challenges.

An overview about the use of research results is shown in Figure 1. Either the results are exploited which means will be protected by IPR e.g. patents, or they are disseminated. The dissemination can be realized in two ways: first as publication in scientific journals or on conferences, second, the research data can be provided in data repository. For both dissemination opportunities the EC follows the Open Access principle.

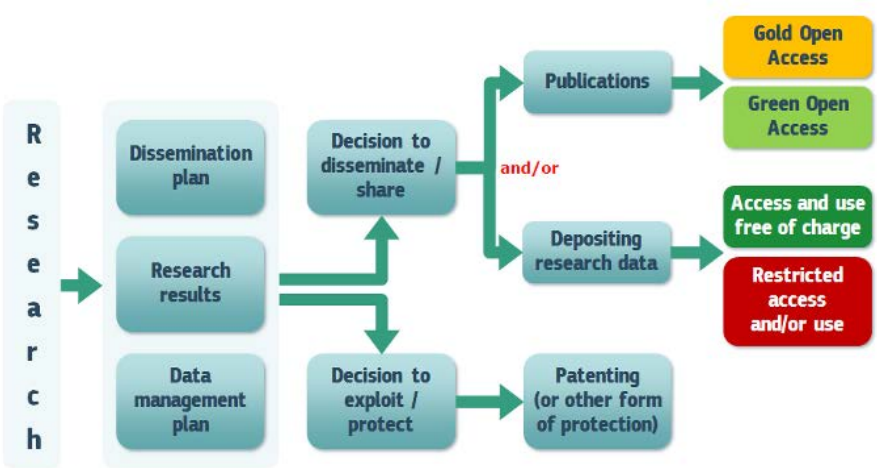

Figure 1: Use of research results [3].

In sense of the OPR pilot, open data means that the research data can typically be accessed, mined, exploited, reproduced and disseminated, free of charge for the user. Research data refers to information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form. The types of data covered by the ORD pilot are:

1. the data needed to validate the results presented in scientific publications including the associated metadata, and
2. any other data (e.g. raw data) including the associated metadata.

The open access to the research data can be denied in cases of:

- results are commercially or industrially exploited,
- incompatibility with confidentiality and security issues – IPR,
- protection of personal data – privacy,
- jeopardise the achievement of the main aim of the action,
- project does not generate and collect any research data,
- other legitimate reason.

To use and provide open access to scientific research data, it should be easily discoverable, accessible, assessable and intelligible, useable beyond the original purpose for which it was collected as well as interoperable to specific quality standards. These qualities will be ensured by the data management plan (DMP). The DMP provides information on and specifies the data the research will generate, how to ensure its curation, preservation and sustainability, what parts of that data will be open and how. Its purpose is to support the data management life cycle and provide an analysis of the main elements of the data management policy of all the datasets that will be generated by the project. Therefore, it rules the data handling and refers to the right to access and reuse digital research data under its terms and conditions during and after research project.

The DMP is not a fixed but rather a living document. It will be matured during the project and will be continuously updated at least for each project review (mid-term and at end of project) and when:

- new data types are acquired,
- changes in consortium policies occur, or
- changes in consortium composition occur.

This report describes the data handling of the DIMENSION project: which data will be created and how as well as if they are made accessible to public and how. First procedures for data management were already described in the DIMENSION description of action (DoA) [4].

1.2 *DIMENSION DMP*

This document is the initial DMP of the DIMENSION project describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved. The plan will be made available on the project website and continuously updated by all project partners according to the project developments and the Horizon 2020 guidelines. Besides the Consortium Agreement, the DMP will be used to manage (amongst other things) the ownership and access to key knowledge of the project by effective data management and IPR protection procedures.

The purpose of this DMP is to provide an analysis of the main elements of the data management policy that will be used by the participants of the DIMENSION project with regard to all the datasets that will be generated by the project. It is a document outlining how research data will be handled during a research project, and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved.

DIMENSION is a project that will create research scientific data in the disciplines of datacentre networks, optical communications and component design as well as fabrication. As DMPs are being enforced only recently by the EC and other major funding bodies worldwide, standardized procedures are not readily available and applicable to research projects covering these disciplines. Nevertheless, DIMENSION will consider the following approaches as far as applicable for providing the open research data:

- data available on the Web (whatever format) under an open licence,
- available as structured data (e.g. Excel instead of a scan of a table)
- use non-proprietary formats (e.g. CSV instead of Excel)
- use URIs to denote things, so that people can point at your stuff
- link data to other data to provide context

To provide the open access to the research data, the DIMENSION data management policy follows 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 DIMENSION consortium will take 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.

In the following sections the DIMENSION data management will be defined. After this introduction, Chapter 2 will summarize the research data which will be generated, collected, processed and stored. Different types of data will be defined and described with respect to its origin, properties, and (re-)use. In Chapter 3, the DIMENSION principles for FAIR (findable, accessible, interoperable, re-usable) data will be defined. In Chapter 4 the financing and responsibility for FAIR data will be briefly stated. Chapter 5 describes the measures which are taken for data security. Chapter 6 and 7 deal with ethical and other aspects regarding open research data. Finally, Chapter 8 summarized the DIMENSION data management.

2 Data Summary

This section gives and overview about the research data which are generated, collected, processed and stored in DIMENSION. 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.

The state of data generated in DIMENSION 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 DIMENSION. Considering the technical disciplines related to DIMENSION 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

In DIMENSION several components and sub-systems will be designed (e.g. laser, laser driver, modulator, modulator driver, test boards,…). For the design special software like Cadence and Altium Designer is used. Design results are schematics and layouts of the components. The digital format of the designs is mostly appropriate (e.g. gds files) to the special software. Therefore, this software is required for data re-use. Since, these softwares are only available with legal licences, they cannot be provided by consortium. Furthermore, the component design is based on a given (semiconductor) technology physical design kit (PDK), e.g. IHP SG25H1 for the ICs, which is provided by a manufacturer, called foundry. The use of the PDK is only possible under the scope of a non-disclosure agreement (NDA). In addition, the DIMENSION designs and also the integration technology are subject of the main objectives as well as for planned IPR.

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 and emulations. These are used to evaluate and estimate the performance and properties of the simulated device. In DIMENSION simulations are used on different levels. For the design of the components and ICs schematic and layout simulations are performed. With regard to the integration technology and packaging HF

simulations are carried out using 3D EM solvers. The operation of the complete transmitter and optical links will be simulated with higher level system simulations. For the simulations and emulations special software is typically used, which need special software licenses for research or even business use. Such software includes for example OPNET, VPI, MATLAB, HFSS, Sonnet and Cadence. The format of simulation results are for a big extent appropriate data set only usable with the simulation software. In rare 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 is not useful to provide the raw simulation data for open access in DIMENSION. It makes more sense to provide the processed simulation results as diagrams and collected as well as explained in a simulation report. There applicable and 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

Measurement data is similar to simulation data. However, measurement data is generated and collected from hardware characterization and not from software investigations. The results come from lab experiments, system environment demonstrators and hardware analysis. The generated data is related to measurements of specific parameters that indicate the systems, sub-system and component performance. Special optical and HF test and measurement equipment is used to collect the results. Similar to simulation data, also measurement data is 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. In most cases the measurement results are screenshot from the measured diagrams and plots.

Therefore, also in case of measurement data open access of raw data is not promising in DIMENSION. Detailed measurement data will be made accessible after publication via measurement reports which also describe the measurement environment, e.g. specific performance parameters, test and measurement equipment, experimental setups, etc. If applicable database-based measurement data sets will be attached as zip archive to the reports.

Publications

The most open and visible data set of DIMENSION will be indeed the publication of research results. Publications contain a summary of results with description of their generation and corresponding condition. 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 OA approach of Horizon 2020 is embraced by DIMENSION, following the guidelines presented by the Commission. We encourage that the project results will be published mainly at fee-based open access scientific journals, following the OA Gold method, due to the high impact associated with certain journals. Indeed, there exist many open access high-impact journals in the disciplines of optical networks and communications, published by IEEE, OSA and Elsevier allowing a variety of publication venues. For this reason costs for publication fees have been foreseen in the consortium budget.

It is anticipated that our researchers will also primarily target the OA Green method in the case of conferences and workshop contributions, since the two OA 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 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). The Open Access Infrastructure for Research in Europe (OpenAIRE) [5] will be explored by our researchers to determine which repository to choose. At TUD the open access repository Qucosa [6] is available which can be used for all publications with TUD contribution.

Project documents and 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 DIMENSION objectives. Reports can be related to design, simulation and measurements and contains the processed data. The openness or confidentiality of the project reports and documents is directly defined in the DoA and CA of the project. Therefore, as long as the documents are not assigned to be confidential, contain any confidential data or the content is not subject of an IPR, they are per se public. Normally, documents and reports are in standard pdf format. Public DIMENSION documents and reports will be made available and accessible on the DIMENSION webpage after their submission and publication.

3 FAIR data

One of the grand challenges of data-intensive science is to facilitate knowledge discovery by assisting humans and machines in their discovery of, access to, integration and analysis of, task-appropriate scientific data and their associated algorithms and workflows [7]. In this regard FAIR data is a set of guiding principles to make data **f**indable, **a**ccessible, **i**nteroperable and **r**e-usable.

3.1 *Findable data*

ORD has to be findable easily, rapidly and identically. Therefore, exact and standard measures have to be used to identify the data sets. This can include the definition and use of naming conventions, search keywords, version numbers, metadata standards and standard data identifiers.

For non-self-explaining data sets the researcher must ensure that sufficient documentation or metadata (i.e. information about the data, e.g. title, author, dates, access rights) is created and maintained to enable the generated research data to be found, used and managed throughout the project lifecycle. Documentation and metadata requirements will differ depending on the discipline and the nature of the specific activity inside the project. For example reports and publication do not need special metadata since this is already included in the document. In contrast to this, a set of the measurement data, e.g. a set of measured diagrams, need additional explanation of the content of the data set. This will be provided by data documentation which includes context for the data and ensures that the data can be understood in the long term. Metadata can be considered as a subset of the overall data documentation. Common types of metadata include:

- descriptive metadata: identifies the resource and enables it to be discovered
- technical metadata: enables a resource to be better managed, and in some cases preserved over time, by capturing information such as creation and modification dates, file formats and access restrictions.

The most widely-used descriptive metadata standard is Dublin Core that works for different kinds of data (not just digital) and across disciplines. It is a simple metadata standard that is commonly used in institutional repositories. Furthermore, in DIMENSION the UK Data Audit Framework Methodology is considered. Both standards are further assessed with the project progress depending on the data types which further arise.

Since currently only publications, documents and reports are subject of ORD, simple but distinct naming convention of data files are used in DIMENSION as a starting point. The naming convention of DIMENSION files with examples is shown in Table 2. For publication the documents are identified by the digital object identifier (DOI).

Table 2: DIMENSION data file naming convention.

Convention	*[time stamp]_DIMENSION_[data type]_[data type postfix]_[version].[file format]*				
Item	Time stamp	Data type	Data type postfix	version	file format
Optional	X		X		
Definition(s)	YYYY_MM_DD	Design Simulation Measurement Publication Document Report Deliverable Presentation …	arbitrary e.g. LDD, Laser, MD, … e.g. D1.2 e.g. ProjectMeeting …	v#.#	According to software: pdf jpg zip gds xls(x) doc(x) ppt(x) …
Examples	2016_07_31	DeliverableReport	D1.2	v0.1 v1.0	As above
Complete	*2016_07_31_DIMENSION_DeliverableReport_D1.2_v1.0.pdf*				

3.2 *Accessible data*

To enable for third parties to mine, exploit, reproduce and disseminate ORD, it has become accessible. On the one hand this means that the data are stored and provided on a platform which can be accessed by interested parties. These platforms can be for example research data repositories. On the other hand, it has to be guaranteed that the available data itself can be opened and processed. In this regard as much as information on the data-related software or instruments have to be provided, e.g. in a separate document or metadata. It is useless to OA data which needs special and appropriate environments to be accessed. Similar to the re-use of the data, also the modality and conditions for the access itself can be defined by a license if there are conflicts with IPR, privacy- or security-related matters.

Generally, all data which has been published can be made accessible. However, as described in Chapter 2, for most of the data types in DIMENSION appropriate formats and special restrictions (e.g. software licences, technology NDAs, …) are present which prohibit free-access to the data. Most of the data sets for OA are documents, reports and probably collected data from simulation and measurements (e.g. results as screenshot or spreadsheet collection). For this reasons, it has been decided in DIMENSION that the use of a special repository is not beneficial at the moment. The current ORD can be circulated well via the existing webpage as downloads and links. However, a special repository, which links publications to research data, will be evaluated and selected when first data is collected during project progress. Several repositories are considered for this purpose

(e.g. re3data.org, Zenodo, OpenAire-CERN). For joint publications with TUD e.g. can be also placed in the Qucosa repository and will be clearly identified with a DOI. In any case the accessibility of data has to be confirmed by the DIMENSION consortium or contributing partners beforehand.

3.3 *Interoperable data*

Interoperable data means that the exchange and reuse of data enabled by OA is possible. Therefore, standard data vocab and formats which are compliant with available open software have to be used for storing and providing the data. OA of data which can only be used with special and restricted software makes no sense. Only in case of interoperability, data exchange between researchers, institutions, organisations, countries is possible and allows the re-combination with different datasets from different origins.

In this regard DIMENSION will use as wherever possible data formats for OA 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. These archives are linked to a (metadata) document, e.g. the measurement report. All documentation and reports are filed in pdf which is widely read- and usable.

3.4 *Re-usable data*

In order to make the ORD re-usable the conditions for the use have to be defined. Modalities and scope of use can be defined by licenses. This framework indicates if there are any restrictions and why. For example there can be embargo times e.g. to publish and patent during or after project. Furthermore, the durations of re-usability is regulated. Basically, it should be ensured that the data is usable beyond the original purpose for which it was collected even long time after the collection. This enables the interdisciplinary use for new developments after the DIMENSION has finished. Therefore, data is provided in a way that judgments can be made about their reliability and the competence of those who created them. Licenses which rule the use of ORD have to be clear and accessible.

Within DIMENSION re-usability of data is enabled that all data which will be provided for OA yet, e.g. documentation and reports, is inherently public (and free of charge). Therefore, no special licenses are established at the moment. Once, special data will be provided for OA also special licenses will be issued. However, although the data will public the DIMENSION project and its consortium members reserve the copyright of the material. Data sets defined as ORD will be available and accessible after the material was published. In all cases, DIMENSION consortium or

contributing partners have to agree beforehand something becomes OA. For re-usability the data will be stored on the webpage or on a repository system when implemented for at least ten years.

4 Allocation of resources

Currently there are no costs to provide FAIR data. For hosting the DIMENSION webpage, which also provides public information, reports and project documents, a yearly fee of <100 € for using the web domain is required. These costs are covered the project budget of the coordinator TUD. Furthermore, costs for publication fees in open access journals have been foreseen in the consortium budget.

Responsible partner for the DMP and the FAIR data is the coordinator TUD.

5 Data security

Research data generated, processed and collected in DIMENSION is stored on computers, clusters and servers at each project partner's premises. The facilities are hosted by the partners themselves and are secured according to actual security guidelines. The data are placed in a back-up storage to be able to restored in case of emergencies. The storage duration is usually 10 years according to the funding rules.

General project data and documentation is also stored in the project SharePoint. Furthermore, public documents are also provided on the project webpage. Both, the SharePoint and webpage, are hosted by the coordinator TUD in house. Data security, backup and recovery is applied and the storage duration is also at least 10 years.

6 Ethical aspects

In the DIMENSION 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 DIMENSION consortium is no subject of the project's data management and ORD.

7 Other issues

There are no further national, funder, sectorial or departmental procedures to be followed for the data management in DIMENSION. For internal data use each partner applies the data management defined in its institution. The DIMENSION documentation is also stored in the project SharePoint, which is accessible to all project partners.

8 Conclusion

To conclude the current version of DMP, a summarized overview about the DIMENSION research data set and its properties is shown in Table 2. As described in the previous Chapters, it is currently only possible to provide public project documents and publications. As a general rule, the DIMENSION consortium has to confirm beforehand the accessibility and re-usability of the data provided for ORD. This DMP represents the current state of data handling in the DIMENSION project. As living document it will be continuously updated at least at the end of each project period or in the cases of new data types are acquired, changes in consortium policies occur, or changes in consortium composition occur.

Table 3: Summary of DIMENSION data sets and management.

Data type	Origin	Purposes	Formats	Re-usability	Open access	Comments
Design	Design software, e.g. Cadence, Altium Designer	Schematics, layouts for fabrication	Appropriate, e.g. gds	With corresponding design software; licenses; NDAs	Not yet	Design PDK subject to NDA
Simulation	Simulation software, e.g. OPNET, VPI, MATLAB, HFSS, Sonnet, Cadence	Estimation/analysis of component or system performance based on schematics, layout	Appropriate formats; database and spreadsheet formats; screenshots / pictures / diagrams	With corresponding simulation software; licenses; zip archives with data sets attached to reports, after publication	Not yet	Special software connate be provided; licenses required
Measurement	Measurement device, setup, lab experiments	Verification/analysis of component or system performance in hardware	Equipment appropriate formats; screenshots / diagram pictures	With special equipment; diagram pictures as zip archives attached to reports; after publication	Not yet	Equipment and appropriate software cannot be provided
Publication	Project results; Writing tools / software, e.g. MS Word or Latex	Presentation of project results	pdf	With pdf readers; golden and green OA; After embargo times of publisher	Yes	Via DIMENSION webpage and/or via repositories, e.g. Qucosa
Documents / reports	Project results/progress; Writing tools / software, e.g. MS Word or Latex	Summarize and report project progress and results	pdf	With pdf reader	Yes	Via DIMENSION webpage

References

[1] Guidelines on Data Management in Horizon 2020; Version 1.0, Dec. 2013.

[2] Guidelines on Data Management in Horizon 2020; Version 3.0, July 2016.

[3] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

[4] DIMENSION project, “Description of Action”, October 2015.

[5] http://www.openaire.eu

[6] http://www.qucosa.de/

[7] FAIR Data Management in Horizon 2020, V3.0, July 2016.

Acronyms

Acronym	Definition
DMP	Data management plan
DoA	Description of action
DOI	Digital Object Identifier
EC	European commission
HF	High frequency
IC	Integrated Circuit
IPR	Intellectual property right
NDA	Non-disclosure agreement
OA	Open access
ORD	Open research data
PDK	Physical design kit
TUD	Technische Universität Dresden
URI	Uniform resource identifier