D8.2

Data Management Plan (DMP)

Project number:	644039
Project acronym:	M3TERA
Project title:	M3TERA – Micromachined terahertz systems – a new heterogeneous integration platform enabling the commercialization of the THz frequency spectrum
Start date of the project:	1st February, 2015
Duration:	36 months
Programme:	H2020-ICT-2014-1

Deliverable type:	Report
Deliverable reference number:	ICT-644039 / D8.2 / 1.0
Work package contributing to the deliverable:	WP8
Due date:	JUL 2015 – M06
Actual submission date:	29th July, 2015

Responsible organisation:	ERICSSON
Editor:	Corinna Kudler (TEC) and Yinggang Li (ERICSSON)
Dissemination level:	Public
Revision:	1.0

Abstract:	The purpose of this DMP is to provide an analysis of the main elements of the data management policy that will be applied with regard to all the datasets to be generated in M3TERA including the process, methodology and policy on data generation/ collection, handling, archiving/ preservation and data application.
Keywords:	Data management policy, research data, datasets, IPR, public availability

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644039.

This work is supported (also) by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0059. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government.

Editor (ordered according to beneficiary numbers)

Corinna Kudler (TEC)

Yinggang Li (ERICSSON)

Contributors

All M3TERA partners contributed to this deliverable

Internal Reviewer (ordered according to beneficiary numbers)

Itziar Maestrojuan (ANTERAL)

Executive Summary

“Data” is defined as materials generated or collected during the course of conducting research. Many variables govern what constitutes “data” and the management of data, and each project has its own culture regarding data. This data management plan (DMP) provides an instruction on how the data generation and publication process is handled within the M3TERA project.

The project consortium has defined the expected generated data types by filling in a questionnaire (see Appendix). The largest amount of data will be generated through performance measurements, but also through simulations considering that research will be part of the M3TERA outcome. This data management plan ensures that the data is documented properly and IPR rules, defined within the project consortium, are respected properly. The M3TERA data will be made available on the project website as well as promoted to interested target groups through various further dissemination channels. This DMP also handles the data preservation including cost management as well as the plan regarding how data will be stored for at least three years after the project end.

Further details can be found in the elaborated sections. Please note that this document is planned to be kept as a living document and is meant to be updated on a regular basis to deepen the strategy and plans of the M3TERA consortium to make project data publically available.

Contents

Chapter 1 Introduction

The M3TERA Data Management Plan (further on referred as DMP) is required for H2020 projects participating in the Open Research Data Pilot and describes the data management life cycle for all data sets that will be generated, collected, and processed by the research project M3TERA. Being more specific, it outlines how research data will be handled, what methodology and standards will be used, whether and how the data will be exploited or made accessible for verification and re-use and how it will be curated and preserved during and even after the M3TERA project is completed. The DMP can be considered as a checklist for the future, as well as a reference for the resource and budget allocations related to the data management.

However, to explain the **reason** why a DMP gets elaborated during the lifespan of a research project, the European Commission's vision is that information already paid for by the public purse should not be paid again each time it is accessed or used. Thus, other European companies should benefit from this already performed research.

To be more specific, *"research data refers to information, in particular facts or numbers, collected to be examined and considered and 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 recording and images. The focus is on research data that is available in digital form."*[1]

The DMP is not a fixed document. It will evolve and gain more precision and substance during the lifespan of the M3TERA project. The first version of the DMP, including information from the first six months of the project, includes the following:

- Data management
  - Data set description
  - Collection/Generation/Documentation of Data and Metadata
  - Intellectual Property Rights
- Accessibility
  - Data access and –sharing
  - Archiving and preservation

However, before this information from all partners gets depicted in a more detailed manner (Chapter 3 and Chapter 4), first the used methodology (Chapter 2) gets shortly described in the following chapter.

[1] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Chapter 2 Methodology

As mentioned in the introduction, a research instrument questionnaire was selected as the best mechanism of collecting partner inputs related to the data management within the M3TERA project. This had the dual aim of first gathering a more detailed understanding of the operations planned during the project and also to raise the awareness of the requirements outlined in the Guidelines on Data Management in Horizon 2020.

The questionnaire has been divided into five main chapters, consisting each one of a series of questions (note: the questionnaire template is attached to Appendix of this document). In total, the questionnaire was designed to be broad enough to include the information required by the European commission and to cover the various roles the partners play within the M3TERA project. As the project is by now within its first months, some information remains undefined at the moment. Therefore, a more detailed and elaborated version of the DMP will be delivered at later stages of the project.

Moreover, the DMP will be updated at least by the mid-term and final review to be able to fine-tune it to the data generated and the uses identified by the consortium.

Chapter 3 Data management

The term 'Data Management' stands for an extensive strategy targeting data availability to target groups within an organized and structured process converted to practice. Before making data available to the public, the data to be published needs to be defined, collected, documented and addressed properly. The following sections define this process within M3TERA and will be led by the following questions:

- **3.1 Data Set** – Which type of data will be generated? Which formats will be chosen and can be reused? Which data volume will the data comprise?
- **3.2 Data Generation and Collection** – How can the data set be described? To whom might be the data useful? How can it be identified as research data?
- **3.3 Data Documentation &** - Does the project data comply with international research standards?
- **3.4 Intellectual Property Rights** - Will the public availability be restricted due to the adherence to Intellectual Property Rights?

3.1 Data Set Description

The project has been generating data during the lifespan of the M3TERA project. The overall volume of the generated data is estimated to reach 5-10 GByte. Approximately one out of four beneficiaries will reuse existing data for M3TERA, whereas the rest will start generating and collecting data from scratch. This generated data will differentiate then from quality data, over characterization data of microsystem components and subsystem performance, provided as graphs and raw data, to data design flow or mixed-signal circuit design data (Virtuoso, Spectre, Avenue, Calibre).

Furthermore, geometrical designs (STEP), simulations, measurements (CSV), and calculation data will be generated. Also Matlab will be used as data generation instrument. Mostly data will be displayed in numbers and/or pictures, geometrical- (*.sm3, *.jpg, *.dxf), Microsoft Excel and Microsoft Word format, and through the use of the EM software (*.hfss).

Moreover, the consortium acknowledged that the chosen formats and used in-house software will enable long-term access to the mentioned data.

3.2 Data Generation and Collection

Data generation and collection is concerened about the project data generated or collected, including its origin, nature and scale, and to whom it might be useful.

Data will mostly be generated by the M3TERA beneficiaries themselves or among the consortium. Therefore, different methodologies come into operation. Almost half of the partners will generate data via research. Others will do different types of measurements and simulations (e.g. microwave, process/device, and other components), or bottom-up/top-down design flow (behavioural, transistor level, circuit synthesis, hand layout, layout synthesis, verification, etc). However, for some partners the exact methodology is unknown yet, but important to use what are commonly used and known. In case the data gets collected and not generated in M3TERA, data originally will come from literature research, internal databases, company internal instrumentation, and through design (e.g. MMIC), simulations and measurements.

The consortium agrees in prospectively seeing the possibility to integrate or reuse the generated data. They further agree that the data will be useful for universities, research organizations, SMEs and scientific publications. Moreover, it might be also beneficial for IP providers and to design companies. Even though the data either includes already information for the use or is nonetheless so transparent to not require information to be read and interpreted, half of the partners mentioned that dedicated software packages, access to the PDK and IFAT design flow and tools are required.

3.3 Data Documentation & Metadata

Data documentation ensures that the given dataset or set of documents will be understood, citied properly and interpreted correctly by everyone.

All partners will document their data in a different way, either logging relevant data, or using dedicated software (EM/EDA), libraries and IP management systems. Others prefer to document it after designing, simulating and measuring the components using also MS office and MATLAB. Almost half of the partners will not use metadata standards, the rest however will use EAD, ISO/IEC, SAML, Cadence and .xml formats.

3.4 Intellectual Property Rights

Even though IPR issues mainly arise during the project lifetime or even after project end due to the dissemination (scientific and non-scientific publications, conferences etc.) and exploitation (licensing, spin-offs etc.) of project results, the M3TERA consortium considered the handling of IPR right from the very beginning, already during the project planning phase. Therefore a Consortium Agreement (CA) clearly states the background, foreground, sideground of each partner and defines rules regarding patents, copyrights, (un-) registered designs and other similar or equivalent forms of statutory protection.

Within the M3TERA project most data will be generated within internal processes at partner level through measurement analysis. Close cooperation within the consortium may lead to joint generation of data, which is clearly handled in terms of IPR issues within the CA.

At this stage of the project, no licenses are required, as the commercial value of the data itself might be low. The reuse of valuable data within M3TERA is covered by the CA and will be depending on hardware and software targets of the consortium.

Furthermore, no third party data is reused in the current project phase. In case third-party data will be reused, confidentiality restrictions might apply in specific cases, which will be analyzed per case in detail.

Project data will be published only after review or publication through scientific publication institutes or after ensuring that data is uncritical in terms of IPR issues. Further, data of commercial value for the project partners might underlie restrictions or face a minor time lag before publication.

In total, within M3TERA, all public data is well discoverable and accessible. However, confidential data is only accessible via internal partner platforms and the provided IT infrastructure solely for the M3TERA consortium as agreed in the Consortium Agreement. As data gets and will be provided in readable text format, the consortium (except Chalmers) agrees that the data is assessable and intelligible. Further, they confirm that as a basis for future scientific research activities the data will be usable beyond the original purpose. Regarding suitable standards it can be finally said that only ANTERAL mentioned that data is interoperable to specific quality standards, whereas the others cannot state a comment at the moment, do not know it or deny this statement.

Chapter 4 Accessibility

While Chapter 3 focuses on the internal project processes before publication including the compliance with the project rules for IPR, Chapter 4 describes how the generated data will become accessible for public (re-) use (Section 4.1) and how the availability will be ensured permanently, whether data needs to be destroyed/retained for any contractual, legal or regulatory purpose as well as how long the data should be preserved, what costs will occur and how they will be covered. (Section 4.2).

4.1 Access and Sharing

Access to and sharing of data helps to advance science and to maximize the research investment. A recent paper[2] reported that when data is shared through an archive, research productivity and often the number of publications increases. Protecting research participants and guarding against disclosure of identities are essential norms in scientific research. Data producers should take efforts to provide effective informed consent statements to respondents, to identify data before deposit when necessary, and to communicate to the archive any additional concerns about confidentiality. With respect to timeliness of data deposit, archival experience has demonstrated that the durability of the data increases and the cost of processing and preservation decreases when data deposits are timely. It is important that data is deposited while the producers are still familiar with the dataset and able to fully transfer their knowledge to the archive.

In particular potential users can find out about generated and existing data most likely through the project's dissemination activities (scientific publications and papers), deliverables, presentations and technical events (conferences, trade shows) etc. During the project lifetime these documents and data will be published on our official project website (www.m3tera.eu) where a broad community has access to the project information. Besides the M3TERA public websites also marketing flyers or the internal project SVN repository will be used as a tool to provide and exchange the requested data.

In principle, the data will be shared within the M3TERA consortium according to our Consortium Agreement (with respect to any IPR issues) via a secured SVN repository as soon as the data is available. To the public community, data will be shared according to the dissemination level of the data via the public project website. Partner Ericsson stated that they will share their data to the public under bilateral agreements but there are no conditions for "open" data generated by them. Besides the SVN and the website, the consortium is also willing to handle requests directly. Public deliverables will be made available as soon as they have been approved by the European Commission.

In this early stage of the project (M06) the consortium does not pursue to get a persistent identifier for the data generated.

4.2 Archiving and Preservation

Generally, the consortium's opinion is that it will not be necessary to destroy any data for contractual, legal, or regulatory purposes. However, as described before, there will be the case that the confidential deliverables will be restricted.

[2] http://deepblue.lib.umich.edu/handle/2027.42/78307

At the moment it cannot be determined if other data should be kept. Along with the project progress, the M3TERA consortium will discuss this further. However, the data generated will serve as basis for future scientific research work and reports on device performance as well as for benchmarking. The M3TERA consortium will use the data also for the development of SiGeBiCMOS circuits, the use of mm-wave suited packages (eWLB), for chip/RF-MEMS interfaces and the RF-MEMS design and modeling. Further foreseeable research will be mmW building practice, future radio systems as well as antenna research. The consortium will also develop a sensing prototype for M3TERA.

With regards to the retention and preservation of the data, M3TERA will retain and/or preserve the produced data at least for three years after the project end. Further, it will be stored in a commodity cloud with usage of internal infrastructure and data bases from the partners or external platforms.

Costs for data storage and archiving will occur, in particular for sever provision (infrastructure) and maintenance (security updates). The coordinator, Technikon, has foreseen appropriate costs in the project budget for the active project time. At a later stage of the project it can be better assessed, if further costs for data storage will occur. These costs will then be covered by the partners with their own resources.

Chapter 5 Summary and conclusion

This data management plan outlines the handling of data generated within the M3TERA project, during and after the project lifetime. As this document will be kept as a living document it will be regularly updated by the consortium. The partners put into write their plans and guarded expectations regarding valuable and publishable data.

A questionnaire on data management issues supported the partners to create awareness for data handling right at the project start. Within the M3TERA consortium qualitative data, characterization data, design data etc. will be generated in different designs like Matlab, Microsoft Excel, EM etc. These data will be valuable for universities, research organizations, SMEs and scientific publications.

The M3TERA consortium is aware of proper data documentation requirements and will rely on each partners' competence in appropriate citation etc. The Consortium Agreement (CA) forms the legal basis in dealing with IPR issues and covers clear rules for dissemination or exploitation of project data. Besides the M3TERA public website, which targets a broad interest group, also marketing flyers or the SVN repository will be used as a tool to provide data. With regards to the retention and preservation of the data, M3TERA partners will retain and/or preserve the produced data for several years, three years after the project end at least.

The M3TERA consortium is convinced that this data management plan ensures that project data will be provided for further use timely, available and in adequate form, taking into account the IPR restrictions of the project.

Chapter 6 List of Abbreviations

Abbreviation	Explanation
CA	*Consortium Agreement*
CSV	*Comma-separated values*
DMP	*Data Management Plan*
EC	*European Commission*
EM	*Electromagnetic*
H2020	*Horizon 2020*
ICT	*Information and Communication Technologies*
IPR	*Intellectual Property Rights*
PDK	*Process Design Kit*

Chapter 7 Appendix – DMP Questionnaire

Data Management Plan (DMP) - Questionnaire

Introduction on Data Management Plan [please read the introduction carefully!]

To explain the **reason**, why performing a DMP is, that the EU Commission's vision is that information already paid for by the public purse should not be paid again each time it is accessed or used. Thus, other European companies should benefit from our research.

The **Data Management Plan (DMP)** is a document outlining how research data will be handled during the M3TERA project and after it is completed.

*"**Research data** refers to information, in particular facts or numbers, collected to be examined and considered and 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 recording and images. The focus is on research data that is available in digital form."*[3]

It should always include clear descriptions and rationale for the access regimes that are foreseen for collected data sets.

A DMP described the data management life cycle for **all data sets** that will be collected, processed or generated by the research 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.

The DMP is **not a fixed document**; it evolves and gains more precision and substance during the lifespan of the project. It should include the following information:

- **Description of Data**
- **Data Collection/Generation**
- **Data management, -documentation & Metadata**
- **Intellectual Property Rights**
- **Accessibility – Data sharing, archiving and preservation**

The first version of DMP is expected to be delivered within the first 6 months of the project and should be compatible with the template provided by the Commission.

More elaborated versions of the DMP can be delivered at later stages of the project. The DMP would need to be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start.

[3] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Questionnaire to collect data from partners - please fill in (see yellow marked fields!)

Description of Data	
Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.	
Will you generate any type of data?	[please fill in]
If yes, what type, format and volume of data? [Data set reference and its name]	[please fill in]
Do your chosen formats and software enable sharing and long-term access to the data?	[please fill in]
Are there any existing data that exist/you can reuse (link/information)?	[please fill in]

Data Management: Collection/Generation/Documentation of Data & Metadata	
Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed. Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.	
Who created/contributed/owns to the data?	[please fill in]
What is the used methodology?	[please fill in]
What is the data's origin (in case it is collected)?	[please fill in]
For whom/end user is the data useful (e.g. university, research organization, SME's, scientific publication)	[please fill in]
Do you see any possibility to integrate or reuse the data in future?	[please fill in]
What information is needed for the data to be read and interpreted in the future?	[please fill in]
How will you capture / create this documentation and metadata?	[please fill in]
What metadata standards will you use and why?	[please fill in]
Research Data Identification	

Discoverable:

Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)

Accessible:

Are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)

Assessable and intelligible:

Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgements can be made about reliability and the competence of those who created them)?

Useable beyond the original purpose for which it was collected

Are the data and associated software produced and/or used in the project usable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?

Interoperable to specific quality standards

Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins?)

Is the data discoverable?	[please fill in]
Is the data accessible?	[please fill in]
Is the data assessable and intelligible?	[please fill in]
Is the data usable beyond the original purpose for which it was collected?	[please fill in]
Is the data interoperable to specific quality standards?	[please fill in]

Intellectual Property Rights	
How will the data be licensed for reuse?	[please fill in]
Are there any restrictions on the reuse of third-party data?	[please fill in]
Will data sharing be postponed / restricted e.g. to publish or seek patents?	[please fill in]

Accessibility – Data sharing, archiving and preservation

[Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.)

Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how

these are planned to be covered.] In case the dataset cannot be shared, the reason for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).]	
How will potential users find out about your data?	[please fill in]
With whom will you share the data, and under what conditions?	[please fill in]
Will you share data via a repository, handle requests directly or use another mechanism?	[please fill in]
When will you make the data available?	[please fill in]
Will you pursue getting a persistent identifier for your data?	[please fill in]
What data must be retained/destroyed for contractual, legal, or regulatory purposes?	[please fill in]
How will you decide what other data to keep?	[please fill in]
What are the foreseeable research uses for the data?	[please fill in]
How long will the data be retained and preserved?	[please fill in]
Where e.g. in which repository or archive will the data be held?	[please fill in]
What costs if any will your selected data repository or archive charge?	[please fill in]
How will these costs be covered?	[please fill in]
Have you costed in time and effort to prepare the data for sharing / preservation?	[please fill in]