



FVLLMONTI

Call: H2020-EU.1.2.2.

Grant Agreement no. **101016776**

Deliverable D6.6 – Data Management Plan

Start date of the project: 1st February 2021

Duration: 60 months

Project Coordinator: Cristell MANEUX - University of Bordeaux

Contact: Cristell MANEUX - cristell.maneux@ims-bordeaux.fr

DOCUMENT CLASSIFICATION

Title	Data Management Plan
Deliverable	D6.6
Estimated Delivery	31/08/2021 (M6+2)
Date of Delivery Foreseen	31/08/2021 (M6+2)
Actual Date of Delivery	31/08/2021 (M6+2)
Authors	Jens Trommer – P6 – NAMLAB GGMBH, Ian O’Connor – P3 – ECL, Georgeta Bordea – P1 – UBx
Approver	Cristell Maneux – P1 – UBx
Work package	WP6
Dissemination	PU
Version	V1.0
Doc ID Code	D6.6_FVLLMONTI_P6-NAMLAB-GGMBH-20210831
Keywords	Data management, Open datasets

DOCUMENT HISTORY

VERSION	PUBLICATION DATE	CHANGE
1.0	28.07.2021	Initial version

DOCUMENT ABSTRACT

The FVLLMONTI Data Management Plan describes datasets generated and published during the duration of the project.

Data Management Plan Information

FVLLMONTI Data management plan

This document describes the datasets generated in the FVLLMONTI project

Funder

European Commission | EC

Grant

Ferroelectric Vertical Low energy Low latency low volume Modules fOr Neural network
Transformers In 3D

Organisations

University of Bordeaux

Researchers

Jens Trommer (orcid:0000-0003-2972-438X), Georgeta Bordea (orcid:0000-0001-9921-8234),
Cristell Maneux (orcid:0000-0001-9125-5372), Ian O'Connor (orcid:)

Datasets

Title: Transformer models for Automatic Speech Recognition (ASR) and Machine Translation (MT)

Template: Horizon 2020

This dataset describes the pre-training parameters of the Transformer language representation model along with the fine-tuning parameters for the tasks of automatic speech recognition and machine translation in English, French and German.

Dataset Description

1.2 Data Summary

1.2.2 What is the purpose of the data collection/generation and its relation to the objectives of the project?

[To share information, To keep on record, To make informed decisions, To combine with other data]

Comment: Application-specific data is collected to validate and assesses the reliability of the 3D neural network compute cube (N2C2) based on vertical nanowire field-effect transistors (VNWFET). This dataset will support one of the main objectives of the project, that is to appraise the impact of wear out failure mechanisms on the whole FVLLMONTI value chain: N2C2, 3D transformer architecture and up to the Automatic Speech Recognition (ASR) and Machine translation (MT) application. Beyond the specific translation application, the final intent is to demonstrate the intrinsic 3D Neural Network architecture robustness.

1.2.3 What types of data will the project generate/collect?

[derived or compiled (e.g., text mining)]

(a) Pre-training and fine-tuning configuration parameters: number of layers (i.e., Transformer blocks), the hidden size, number of self-attention heads, total number of parameters. (b) Corpus information: languages, number of words, number of sentences (c) Test results measured in Word Error Rate (WER) and Bilingual Evaluation Understudy (BLEU) score. (d) Trained model for inference on sample data (e) Sample test corpus

1.2.4 What formats of data will the project generate/collect?

[Models - 3D, statistical, Software - Java, C, Python]

(a) Parameter configuration (.txt, .json files) (b) Scripts (.sh files) (c) Python code (.py files) (d) Pre-trained models

1.2.6 What is the origin of the data?

[Secondary data]

1.2.7 What is the expected size of the data?

GB (gigabyte)

Comment: The largest files included in the dataset are the pre-trained models which are typically around 500MB.

1.2.8 To whom might it be useful ('data utility')?

[Researchers, Research communities, Industry]

Academia and industry for reproduction, utilisation for further applications, and to raise the quality of their final products for the end user.

2.3 Reusable Data

2.3.1 Will you re-use any existing data and how?

Yes

<https://github.com/espnet/espnet> <https://paperswithcode.com/dataset/must-c>

[To compare and combine with other data, To follow-up research on a specific area]

3.4.2 Making data findable, including provisions for metadata

3.4.2.1 Will you use metadata to describe the data?

No

3.4.2.3 Will your metadata use standardised vocabularies?

No

3.4.2.7 Will you use naming conventions for your data?

Yes

3.4.2.8 Please provide more details and examples on used naming conversions

Data will be easily recognized using the following format

YYMMDD_[Model]_[Corpus]_[Language], where YYMMDD represents the inverted date of the day.

3.4.2.9 Will you provide clear version numbers for your data?

Yes

Comment: Versioning will be automatically done using GitHub.

3.4.2.10 Will you provide persistent identifiers for your data?

Yes

Comment: These collections will be linked to scientific articles, conference proceedings, reports, and other sources to be published. We will make use of persistent and unique Digital Object Identifiers (DOI) via the data storage facility.

3.4.2.11 Persistent identifiers

DOI

3.4.2.12 Will you provide searchable metadata for your data?

Yes

Comment: Keywords will be assigned to the dataset.

3.4.2.13 What services will you use to provide searchable metadata?

Registry/Catalogue

OpenAIRE

3.4.2.15 Will you use standardised formats for your data?

No

3.4.2.18 Are the file formats you will use open?

Yes

Comment: All data will be made available in open data formats.

3.4.2.21 Do supported open-source tools exist for accessing the data?

Yes

Zenodo

3.4.2.23 Will you provide metadata describing the quality of the data?

No

3.4.3 Making data openly accessible

3.4.3.1 Are there ethical or legal issues that can impact sharing the data?

No

3.4.3.3 Will your data be openly accessible?

all

3.4.3.5 How will the data be made available?

[Project website, Repository of Archive]

Zenodo

GitHub

3.4.3.7 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

Comment: Storage Area Network with password protection, RAID structure, and daily tape backups.

3.4.3.10 Are there any methods or tools required to access the data?

Yes

3.4.3.11 Please provide information about the method(s) needed to access the data

ESPNET

<https://github.com/espnet/espnet>

3.4.3.13 Will you also make auxiliary data that may be of interest to researchers available?
after publication

3.4.4 Making data interoperable

3.4.4.1 Will you use a controlled vocabulary for your data?

No

3.4.4.2 Will you provide a mapping to more commonly used ontologies?

No

3.4.5 Increase data reuse

3.4.5.2 When do you plan to make your data available for reuse?

after article publication

3.4.5.5 What internationally recognised licence will you use for your data?

Creative Commons Attribution 4.0 International

3.4.5.6 Do you have documented procedures for quality assurance of your data?

No

3.4.5.8 Will you provide any support for data reuse?

Yes

3.4.5.9 How long do you intend to support data reuse?

Up to 5 years

4.5 Allocation of resources

4.5.2 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

[Use of institution infrastructure]

Comment: Through usage of available, free infrastructure, additional costs are avoided.

4.5.3 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

Comment: Responsibilities for data management are taken over by the project manager.

4.5.4 Identify the people or roles that will be responsible for the management of the project data

Georgeta Bordea (orcid:0000-0001-9921-8234)

4.5.5 How do you intend to ensure data reuse after your project finishes?

[Institutional archive]

Comment: Repository of University & Zenodo

5.6 Data Security

5.6.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.7 Ethical aspects

6.7.2 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.7.3 What are the methods used for processing sensitive/personal data?

[Not available]

7.8 Other

7.8.2 Do you make use of other procedures for data management?

No

Comment: Currently we don't use any specific national, funder, sectorial, or departmental procedures for data management.

Title: Vertical Nanowire Platforms Process and Technology Files

Template: Horizon 2020

This dataset describes the basic technologies of the two vertical nanowire platforms explored in FVLLMONTI (junctionless and polarity controllable). It comprises an early process development kit (PDK), layout files and electrical as well as physical characterization data.

Dataset Description

1.2 Data Summary

1.2.2 What is the purpose of the data collection/generation and its relation to the objectives of the project?

["To obtain information", "To keep on record", "To make informed decisions", "To develop a product", "To improve a product", "To combine with other data"]

Comment: All the experimental data is collected to keep in progress and to observe the improvement of the junctionless (JL) and polarity-controllable (PC) vertical nanowire field-effect transistors (VNWFET) on each step of fabrication and characterization. This dataset will not only support the first objective of the project also give very fundamental and crucial results for the entire FVLLMONTI. This data storage will be a guide for the 3D integrated circuit design.

1.2.3 What types of data will the project generate/collect?

["sample or specimen data", "experimental (e.g., energy delay product)"]

(a) layout files: Layouts used for fabrication (laser lithography, e-beam lithography) of VNWFETs and test structures, (b) process parameters: this data will include key process parameters of fabricated devices on each step of fabrication process. (c) Physical Characterization data: this files will contain SEM/TEM images of nanowires, ferroelectric layers and other fabrication step of VNWFETs, (d) Electrical characterization data: I-V characteristics of fabricated VNWFETs.

1.2.4 What formats of data will the project generate/collect?

[".txt files", "PDF", "RTF", "gif", "tiff", "png", "mpeg", "mp4", "Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language)", "Multimedia - jpg / jpeg, gif, tiff, png,

mpeg, mp4, QuickTime", "Instrument specific formats - Olympus Confocal Microscope Data Format, Carl Zeiss"]

(a) layout files (.gds2 files) (b) process description files (.docx . pdf) (c) images (.png, .jpeg .tiff files) (d) raw data and graphs (.opj, .xls .txt files)

1.2.6 What is the origin of the data?

["Primary data"]

1.2.7 What is the expected size of the data?

GB (gigabyte)

Comment: The largest files included in the dataset are the .gds2 layout files and the raw images from physical characterization. A selected will be provided keeping the data amount in the GB range.

1.2.8 To whom might it be useful ('data utility')?

["Researchers", "Research communities", "Education", "Industry"]

Industry and researchers will utilize the data for reproduction and further applications. Results can be also used for educational purpose.

2.3 Reusable Data

2.3.1 Will you re-use any existing data and how?

No

3.4.2 Making data findable, including provisions for metadata

3.4.2.1 Will you use metadata to describe the data?

No

3.4.2.3 Will your metadata use standardised vocabularies?

No

3.4.2.7 Will you use naming conventions for your data?

Yes

3.4.2.8 Please provide more details and examples on used naming conversions

Data will be easily recognized using the following format

YYMMDD_[Technology]_[Datatype]_[Description], where YYMMDD represents the inverted date of the day. [Technology] should give an indicator what transistor type is used. [Datatype] specifies if it is a layout, process or characterization data, [Description] gives the opportunity to add some specifics to help finding the data.

YYMMDD_[Technology]_[Datatype]_[Description]

3.4.2.9 Will you provide clear version numbers for your data?

No

Comment: Experimental characterization data is unambiguously linked to a certain sample, so no versions are required.

3.4.2.10 Will you provide persistent identifiers for your data?

Yes

Comment: These collections will be linked to scientific articles, conference proceedings, reports, and other sources to be published. We will make use of persistent and unique Digital Object Identifiers (DOI) via the data storage facility.

3.4.2.11 Persistent identifiers

DOI

3.4.2.12 Will you provide searchable metadata for your data?

No

Comment: Keywords will be assigned to the dataset.

3.4.2.15 Will you use standardised formats for your data?

No

3.4.2.18 Are the file formats you will use open?

Yes

Comment: All data will be made available in open data formats.

3.4.2.21 Do supported open-source tools exist for accessing the data?

Yes

Zenodo

Comment: It helps data management by giving DOI to various kind of datasheet.

3.4.2.23 Will you provide metadata describing the quality of the data?

No

3.4.3 Making data openly accessible

3.4.3.1 Are there ethical or legal issues that can impact sharing the data?

No

Comment: There is no personal data, biological data or confidential data protected by intellectual property rights collected in this dataset.

3.4.3.3 Will your data be openly accessible?

some

3.4.3.5 How will the data be made available?

["Project website", "Repository of Archive"]

Zenodo

GitHub

3.4.3.7 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

Comment: Storage Area Network with password protection, RAID structure, and daily tape backups.

3.4.3.10 Are there any methods or tools required to access the data?

Yes

3.4.3.11 Please provide information about the method(s) needed to access the data

ESPNET

<https://github.com/espnet/espnet>

3.4.3.13 Will you also make auxiliary data that may be of interest to researchers available?

never

3.4.4 Making data interoperable

3.4.4.1 Will you use a controlled vocabulary for your data?

No

3.4.4.2 Will you provide a mapping to more commonly used ontologies?

No

3.4.5 Increase data reuse

3.4.5.2 When do you plan to make your data available for reuse?

end of project

3.4.5.5 What internationally recognised licence will you use for your data?

Creative Commons Attribution 4.0 International

3.4.5.6 Do you have documented procedures for quality assurance of your data?

No

3.4.5.8 Will you provide any support for data reuse?

Yes

Comment: All people involved in gathering the data can be contacted by e-mail to support data reuse.

3.4.5.9 How long do you intend to support data reuse?

Up to 5 years

4.5 Allocation of resources

4.5.2 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

["Use of institution infrastructure"]

Comment: Through usage of available, free infrastructure, additional costs are avoided.

4.5.3 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

Comment: Responsibilities for data management are taken over by the project manager.

4.5.4 Identify the people or roles that will be responsible for the management of the project data

Georgeta Bordea (orcid:0000-0001-9921-8234)

4.5.5 How do you intend to ensure data reuse after your project finishes?

["Institutional archive"]

Comment: Repository of University & Zenodo

5.6 Data Security

5.6.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.7 Ethical aspects

6.7.2 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.7.3 What are the methods used for processing sensitive/personal data?

["Not available"]

7.8 Other

7.8.2 Do you make use of other procedures for data management?

No

Comment: Currently we don't use any specific national, funder, sectorial, or departmental procedures for data management.

Title: Logic standard cell library based on FVLLMONTI technology

Template: Horizon 2020

This dataset describes, through multiple representations (schematics, sticks diagrams, vsticks (vertical sticks) diagrams, Pareto front data, sized structures: 3D-rendered + GDSII files), elementary (1-bit) logic circuits for typical boolean operations (INV1, NAND2, NOR2, XOR2, SRAM) for CMOS-like and PTL design styles and for combinations of FVLLMONTI technological flavors (1-2-3 stacked gates, chemical-/electrostatic-doping, absence or presence of ferroelectric layer in gate stack).

Dataset Description

1.2 Data Summary

1.2.2 What is the purpose of the data collection/generation and its relation to the objectives of the project?

["To share information", "To keep on record", "To make informed decisions", "To combine with other data"]

Comment: This data is collected to validate VNWFET-based logic circuit designs with a clear identification of technological hypotheses and design style. Validation is intended both in terms of technological feasibility (DTCO - particularly using physical design related representations) and in terms of their use in N2C2 architectures to achieve target performance metrics (STCO - particularly the Pareto-front data).

1.2.3 What types of data will the project generate/collect?

["simulation (e.g., system simulation)", "derived or compiled (e.g., text mining)"]

a) reference circuit schematics using clear symbols to indicate the type and technology of VNWFET used. b) physical design related representations of logic cells such as sticks diagrams, vertical (vsticks) sticks diagrams, GDSII, 3D-representations. c) Pareto-front data extracted from actual simulations using compact models and parasitic extractions illustrating quantitative tradeoffs between delay, energy consumption, volume / footprint, reliability

1.2.4 What formats of data will the project generate/collect?

["Text files - MS Word docs, .txt files, PDF, RTF, XML (Extensible Markup Language)", "Numerical - SPSS, Stata, Excel", "Multimedia - jpg / jpeg, gif, tiff, png, mpeg, mp4, QuickTime", "Models - 3D, statistical", "Other"]

a) Text files (.docx, .pdf) to document circuits. b) Spreadsheet files (.xlsx) to hold simulation results, Pareto-front data. c) Multimedia files (images - .png) to illustrate sticks and vsticks diagrams, 3D-representations

1.2.6 What is the origin of the data?

["Primary data", "Secondary data"]

1.2.7 What is the expected size of the data?

GB (gigabyte)

Comment: A large number of files is expected due to the multiple representations of elementary logic cells, as well as the multiple technological variants and design styles available for a given logic operation.

1.2.8 To whom might it be useful ('data utility')?

["Researchers", "Research communities", "Decision makers", "Education", "Industry"]

Academia and industry for teaching/training, reproduction, utilisation for further applications, and to raise the quality of their final products for the end user.

2.3 Reusable Data

2.3.1 Will you re-use any existing data and how?

No

3.4.2 Making data findable, including provisions for metadata

3.4.2.1 Will you use metadata to describe the data?

No

3.4.2.3 Will your metadata use standardised vocabularies?

No

3.4.2.7 Will you use naming conventions for your data?

Yes

3.4.2.8 Please provide more details and examples on used naming conversions

Data will be organized according to the following convention:

[LogicFunction]_[Datawidth]_[DesignStyle]_[TechnologyVariant]_v[YYYYMMDD] where: LogicFunction represents the Boolean operation (e.g. NAND, NOR, XOR, MAJ ...); DataWidth represents the number of bits in the operands (1bit, 2bit, 4bit ...); DesignStyle represents the design approach (e.g. ComplementaryStatic, DynamicN, DynamicP, PTL ...); TechnologyVariant represents the technology options (e.g. GateStack1, GateStack2, Ambipolar, Ferro ...)

3.4.2.9 Will you provide clear version numbers for your data?

Yes

Comment: Versioning will be automatically done using GitHub.

3.4.2.10 Will you provide persistent identifiers for your data?

Yes

Comment: These collections will be linked to scientific articles, conference proceedings, reports, and other sources to be published. We will make use of persistent and unique Digital Object Identifiers (DOI) via the data storage facility.

3.4.2.11 Persistent identifiers

DOI

3.4.2.12 Will you provide searchable metadata for your data?

Yes

Comment: Keywords will be assigned to the dataset.

3.4.2.13 What services will you use to provide searchable metadata?

Registry/Catalogue

OpenAIRE

3.4.2.15 Will you use standardised formats for your data?

No

3.4.2.18 Please describe the formats you plan to store your data in, including any URLs to documentation.

.docx, .pdf, .xlsx, .gdsii, .png

Commonly used format

3.4.2.18 Are the file formats you will use open?

Yes

3.4.2.21 Do supported open-source tools exist for accessing the data?

Yes

Comment: All data can be accessed by open-source tools. The Cadence database containing schematics and simulation results will be duplicated to open-source documentation (documents for schematics, .csv files for simulation data)

3.4.2.23 Will you provide metadata describing the quality of the data?

No

3.4.3 Making data openly accessible

3.4.3.1 Are there ethical or legal issues that can impact sharing the data?

No

3.4.3.3 Will your data be openly accessible?

some

3.4.3.5 How will the data be made available?

["Project website", "Repository of Archive"]

Github

3.4.3.7 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

Comment: Storage Area Network with password protection, RAID structure, and daily tape backups.

3.4.3.10 Are there any methods or tools required to access the data?

No

3.4.3.13 Will you also make auxiliary data that may be of interest to researchers available?
after publication

3.4.4 Making data interoperable

3.4.4.1 Will you use a controlled vocabulary for your data?

No

3.4.4.2 Will you provide a mapping to more commonly used ontologies?

No

3.4.5 Increase data reuse

3.4.5.2 When do you plan to make your data available for reuse?
after article publication

3.4.5.5 What internationally recognised licence will you use for your data?

Creative Commons Attribution 4.0 International

3.4.5.6 Do you have documented procedures for quality assurance of your data?

No

3.4.5.8 Will you provide any support for data reuse?

Yes

3.4.5.9 How long do you intend to support data reuse?

Up to 5 years

4.5 Allocation of resources

4.5.2 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

["Use of institution infrastructure"]

Comment: Through usage of available, free infrastructure, additional costs are avoided.

4.5.3 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

Comment: Responsibilities for data management are with the WP4 leader.

4.5.4 Identify the people or roles that will be responsible for the management of the project data

Ian O'Connor (orcid:0000-0002-6238-9600)

4.5.5 How do you intend to ensure data reuse after your project finishes?

["Institutional archive"]

Comment: Repository of University.

5.6 Data Security

5.6.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.7 Ethical aspects

6.7.2 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.7.3 What are the methods used for processing sensitive/personal data?

["Not available"]

7.8 Other

7.8.2 Do you make use of other procedures for data management?

No

Comment: Currently we don't use any specific national, funder, sectorial, or departmental procedures for data management.