D1.2: Data Management Plan (DMP)

Revision: v.1.0

Work package	WP 1
Task	Task 1.5
Due date	31/06/2023
Submission date	05/07/2023
Deliverable lead	DLR
Version	1.0
Authors	Tomaso de Cola (DLR)
Reviewers	Jean-Baptiste Milon (MAR)
Abstract	This deliverable deals with the definition, generation, and usage of data throughout the project, hence providing preliminary guidelines and best-practices.
Keywords	Data management plan, datasets, FAIR

Document Revision History

Version	Date	Description of change	List of contributor(s)
0.0	20/05/2023	ToC and organisation of the content	Tomaso de Cola (DLR)
0.1	12/06/2023	DMP questionnaire inputs	Tomaso de Cola (DLR)
0.2	28/06/2023	Consolidation of the contents	Tomaso de Cola (DLR)

0.3	03/07/2023	Finalisation of the deliverable for review	Tomaso de Cola (DLR)
1.0.F	07/07/2023	FInal version for submission	Tomaso de Cola (DLR)

DISCLAIMER

5G-STARDUST (*Satellite and Terrestrial Access for Distributed, Ubiquitous, and Smart Telecommunications*) project has received funding from the **Smart Networks and Services Joint Undertaking (SNS JU)** under the European Union's **Horizon Europe research and innovation programme** under Grant Agreement No 101096573.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

COPYRIGHT NOTICE

EXECUTIVE SUMMARY

The Data Management Plan (DMP) is a living document that identifies which data is produced during the 5G-STARDUST project, who owns that data, how it will be documented, how it will be preserved, and with whom and under what form it will be shared. Most of the data produced during the project will correspond to documentation, source-code, mathematical proofs and experimental results. Most of the documentation that will be generated during the project has already been identified in the Description of Action (DoA). To this end, the present document (deliverable D1.2) provides the first iteration on the DMP on the basis of the consortium understanding of data planned to be generated and handled to be managed in the rest of the project's life. Since the generation of new data or update of this deliverable will be in any case in order, the DMP will be kept as living document for entire duration of the project, publicly accessible, and periodically subject to review and revision.

The data generated during the project will be owned by the partners which have contributed to produce that data. The extent up to which this data will be made available and which restrictions will be imposed on its reuse will be decided on a case-by-case basis by the owners of the data. The partners will comply with the FAIR (findable, accessible, interoperable and reusable) guidelines of the HE programme, which state that data will be made as available as possible, so long that does not negatively affect the commercial advantage of the partners.

The data will be shared among partners using internal repositories or through direct communication; and with the public through the project's website or public repositories. Finally, the data will be preserved from 3 and up to 30 years after the end of the project at the partners' repositories and cloud infrastructures, according to each partner internal policy.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

3GPP	3rd Generation Public Partnership
AI	Artificial Intelligence
API	Application Programme Interface
AW2S	Advanced Wireless Systems
CA	Consortium Agreement
CDR	Call Detail Record
CMS	Communication Management System
CTTC	Centre Tecnològic de Telecomunicacions de Catalunya
DCI	Downlink Control Information
DLR	Deutsches Zentrum für Luft- und Raumfhart
DMP	Data Management Plan
DoA	Description of Action
EARFCN	E-UTRA Absolute Radio Frequency Channel Number
eMBB	Enhanced Mobile Broadband
FAIR	Findable, Accessible, Interoperable, and Reusable
GDPR	General Data Protection Regulation
HTML	HyperText Markup Language
IEEE	Institute of Electrical and Electronics Engineers
IP	IP Address

IPR	Intellectual Property Right
ISP	Internet Service Provider
MAR	Martel Innovate
mMTC	Massive Machine-Type Communications
MNO	Mobile Network Operator
OSI	Open Systems Interconnection
PHP	Hypertext Preprocessor
RBS	Radio Base Station
SFR	Société française du radiotéléphone
SME	Small Medium Enterprise
SRS	Softrware Radio Systems
uRLLC	Ultra Reliable and Low Latency Communications
WP	Work Package

1   INTRODUCTION

The objective of the 5G-STARDUST project is to explore the integration opportunity arising from converging non-terrestrial networks with the terrestrial counterpart under the 5G-Advanced and then possibly influence the 6G development roadmap according to the results and demonstrations carried out in the course of the project. In this light, data will be generated in the course of the project and important dissemination will be carried out to highlight the main achievements of the project. Not less relevant, the project activities will also build on datasets available from the some of the consortium partners to be exploited for running AI-based optimisation frameworks applied to different aspect of radio and network interfaces. These datasets will be re-elaborated towards the objectives of these optimisation frameworks and later made available in dedicated deliverables to make them reusable from the interested audience in other research projects or overall in other relevant and applicable initiatives.
The overall work programme is defined as follows:
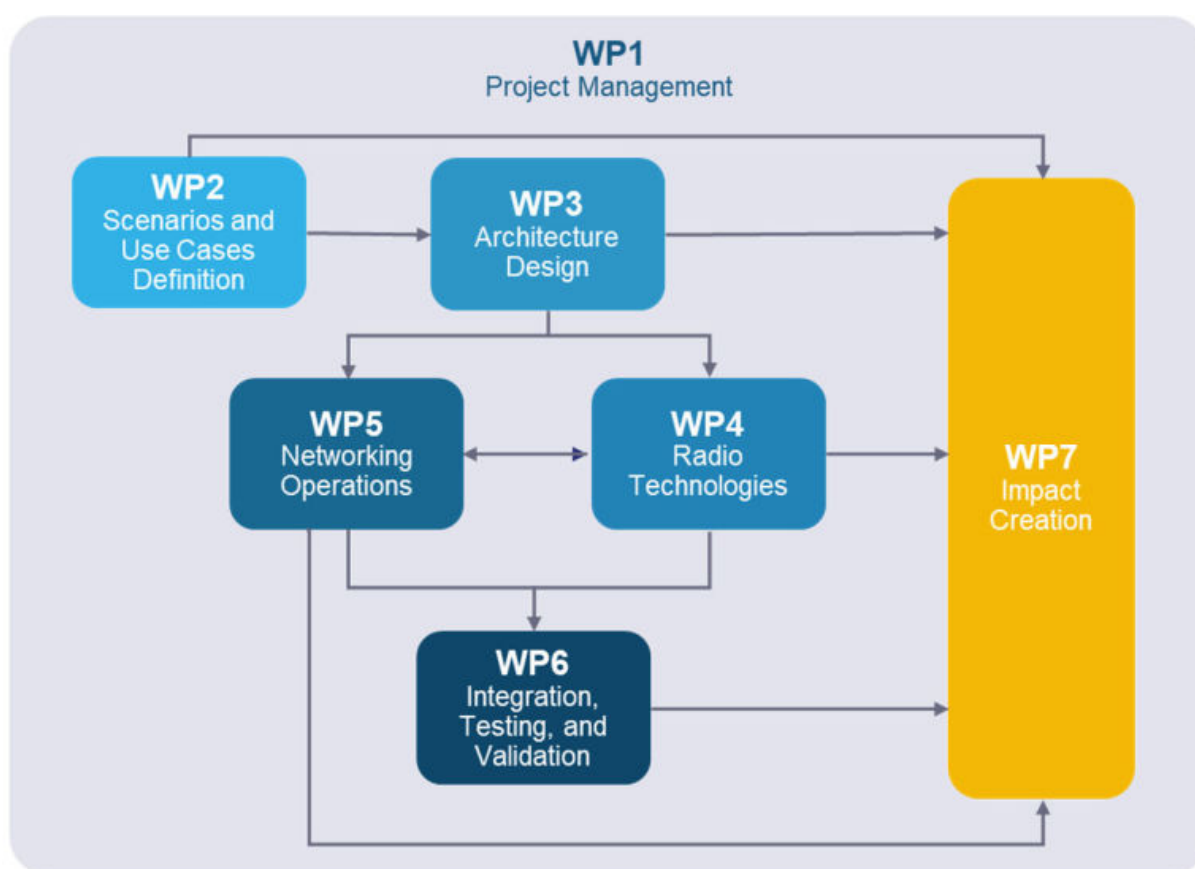
Figure 1 5G-STARDUST Work-programme

The main goal of this document is to identify the datasets (or data collections in a broad sense) that will be produced during the 5G-STARDUST project, research on which standards might be employed to potentiate their reusability and analyse up to which extent they might be disseminated. In this respect, data will be generated and used at the technical development phase (WP4, 5, and 6) and then in the impact creation phase (WP7).
More concretely, this plan describes 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 both during the research project and after it 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. It should be noted that the Data Management Plan (DMP) is not a fixed document: it evolves and gains more precision and substance during the lifespan of the project.

The remaining of the DMP is organised as follows. Chapter 2 includes global considerations about the methodology used to prepare this and future versions of the DMP, the type of datasets that will be generated during the project, which standards will be followed, how they will be documented, preserved and shared, among others. Chapter 3 will consider the generated datasets in more detail, providing more specific features for each, such as who owns each dataset, what it will be used for, and how it will be disseminated and stored. Chapter 4 covers Data Management specificities under the Horizon Europe framework and Chapter 5 is devoted to the underlining the overall publishing framework and related methodologies in terms of tools and internal regulations. Brief notes about Ethics management are given in Section 6, while, Chapter 7 concludes the Data Management Plan by drawing the preliminary conclusions and work ahead.

2 METHODOLOGY AND DATA DESCRIPTION

2.1 METHODOLOGY

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

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

A DMP describes 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.

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 will need to be updated 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.
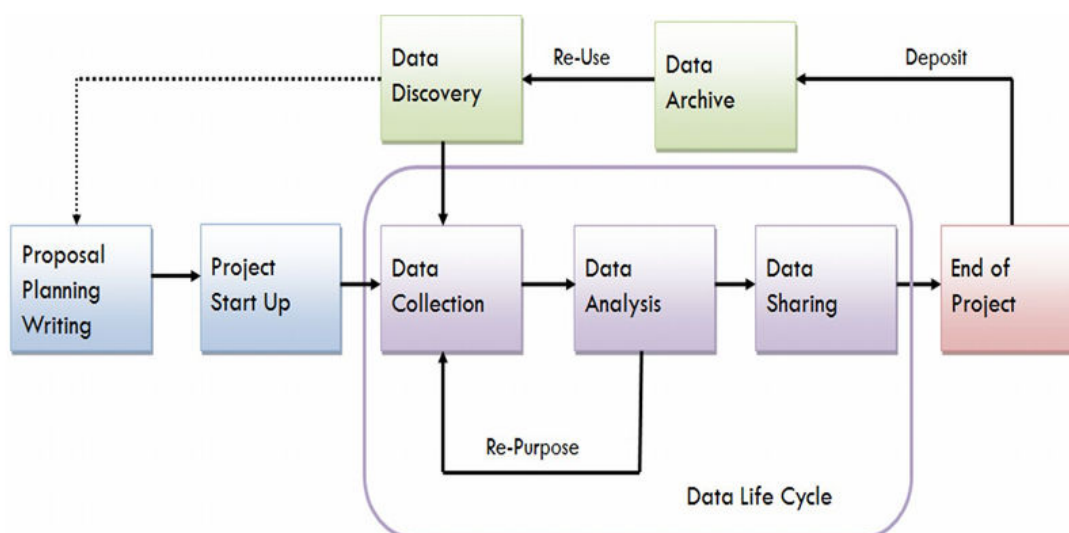

Figure 2 Data Management Lifecycle

Given the heterogeneity of data possibly used by partners (as heritage of previous or other ongoing activities) or generated during this very same project, the preparation of the current Data Management Plan (DMP) has been carried out by distributing a dedicated questionnaire to all partners, in order to get insightful feedback about the current categories:

- Data sets

- Intellectual Property Rights

- Access and Sharing

- Archiving and Preservation

- Documentation and Metadata

The objective of the questionnaire is to identify the current practice of each beneficiary of the project grant and possibly sort out any possible data access or usage restrictions because of confidentiality reasons. This questionnaire being distributed in the first months of the projects, quite some uncertainty still lies in the actual flavour and format of data to be used and generates, whereby future releases of this document will report a more up-to-date status quo about the data management in the project. More importantly, specific datasets owned by consortium partners will be made available for further processing in the context of WP4 and WP5, which will be subject to reporting in dedicated deliverables due end of year 2023, and therefore not specifically addressed here.

Based on the provided information, the following sections and then with more detail Section 3 and 4 of this document have been filled. More importantly a table summarising the data ownership and to which extent corresponding data could be made publicly available is also reported later in this document.

Last but not the least, the reference questionnaire used for this first DMP round is also reported in a dedicated appendix, at the end of the present document.

2.2 DATA COLLECTIONS

Data planned to be collected in the course of the project are of various nature because of the fact data processing and generation is happening in different phases of the project, namely:

- Data generation and processing as output of the functional modules envisaged in the architecture design and then later subject to development and validation

- Data generation as output of the validation process (through simulations and testbed)

- Data generation from the experimental demonstrators, which are collected to demonstrate the effectiveness and the impact of the proposed concepts.

- Datasets provided by consortium partners for use in AI-based procedures (i.e., data driven optimisation, subject to the necessary re-elaboration where needed.

- Data generated by the promotional and dissemination activities

2.3 INTELLECTUAL PROPERTY RIGHTS (IPRS)

The data generated throughout the project will be owned by the partners that have contributed to its creation. The licensing of the data will be agreed upon on a case-by-case manner by the partners involved in each WP, and this document will be updated accordingly.

Open-source libraries that will be used to support the implementations in this project may impose limitations on how the generated system may be distributed.

Second, restrictions on usage might be put in place for code derived from privately owned libraries, which will be identified in the course of the project and then later reported in newer version of the DMP.

2.4 ACCESS AND SHARING

The accessing and sharing of data is firstly ruled by two documents: The Consortium Agreement (CA), which stipulates under which conditions transmitted information between the project partners is deemed confidential and must not be further disseminated; and the Description of Action (DoA) which stipulates the dissemination level of each deliverable. Moreover, the project consortium will comply with the FAIR (findable, accessible, interoperable and reusable) guidelines first introduced in the context of the H2020 programme and the further reaffirmed for the current Horizon Europe programme.

The data necessary to successfully complete the project Work Packages (WPs) will be shared without any restrictions amongst the WP partners either via internal repositories (i.e. internal project sharepoint provided by DLR and other repository provided by MAR in the context of dissemination activities) or direct communication (i.e. emails with attached documentation). Moreover, public data will also be made available at the project's website, Users will be made aware of this data primarily through research publications, patent applications, dissemination activities, invited talks, social networks, the project website (https://www.5g-stardust.eu/), the project sharepoint[1] (https://teamsites-extranet.dlr.de/sites/SNS-2022-STREAM-A-01-02/SitePages/Home.aspx) and other dedicated repositories to be considered in the course of the project in case a need arises. Data will be made available to the project consortium as soon as it is available; to standardisation bodies when required; and to the public at the due date of the derivable. In the case of a research publication based partly on the contents of a published deliverable, it will be made available once the paper gets notification of acceptance. If access to confidential data is necessary by the public, restrictive measures will be put in place.

2.5 ARCHIVING AND PRESERVATION

The project consortium will mostly keep the produced source-code, the accompanying documentation and the experimental results after the end of the project. These will be used for further research that push the state of the art in the core areas, targeted by 5G-STARDUST, of integration of satellite and terrestrial 5G networks, gNB implementation in space, unified

---

[1] The project sharepoint access is currently limited to the projects' users. Availability of certain data can be provided to externals upon agreement

radio interfaces, and optimised radio and network layer performance (through the use of specialised AI-based algorithms).

The generated data will be stored by each partner for a minimum of 3 years beyond the end of the project. The costs associated with storing these data during the course of the project correspond to the server provisioning and maintenance and have been accounted for in the project overheads.

2.6 DOCUMENTATION AND METADATA

Documentation will be provided in formats such as text files, source-code comments, man files, Word documents, Wiki and Markdown. It will be comprised of specification descriptions, code examples of API/library/executable invocations and testbeds. More details about the actual format of these will be provided in later versions of this document.

3   DATA COLLECTION

3.1 CONSORTIUM OWNED DATASETS

The project consortium will take advantage of the datasets later provided by HSP and ORA, towards data-driven optimisation of radio-resource management and networking operations, as addressed in WP4 and WP5. At the time of this deliverable submission, it was not possible to provide more information about these two datasets, since the aforementioned WPs have not yet started and therefore the nature and the characterisation of the datasets to be provided is not yet fully defined nor agreed. This process will be carried out in the second half of year 2023 (i.e. the second half of the first year of the project) and then reported in the two dedicated deliverables [1] and [2], whose submission is planned for the end of the year.

3.2 PUBLIC DATASETS

The consortium has not yet finalised the list of public datasets to be possibly used in the course of the project, but some candidates have been already identified (though final decision to be then taken in the first months of activity of WP4 and 5, i.e. by the end of year 2023), and shortly summarised in the following.

3.2.1   Dataset1: ADOPNET Dataset

Source: https://www-adopnet.irisa.fr/data-description/

Description: The repository consists of the downlink control information (DCI) messages collected in a commercial three-sector eNB of the SFR mobile network operator (MNO) in the city center of Rennes, France. We performed one-week measurement campaign between Saturday, February 19, and Friday, February 25, 2022. The MNO uses four different frequency bands in the monitored sector:

1. EARFCN 6300: 800 MHz band and a bandwidth of 10 MHz

2. EARFCN 1501: 1800 MHZ band and a bandwidth of 20 MHz

3. EARFCN 78: 2100 MHZ band and a bandwidth of 15 MHz

4. EARFCN 2825: 2600 MHZ band and a bandwidth of 15 MHz

For each day and each frequency band, the data is stored in a CSV file named TraceDATE_sfr_earfcnXX.csv, where DATE represents the date in yyyymmdd format and XX represents the EARFCN of the frequency band. In the CSV file, each row represents a DCI message, and the columns include the following features:

* Time: Unix timestamp in milliseconds

* rnti: RNTI associated with the DCI

* SFN: system frame number

* format: DCI format (0: format0, 2: format1A, 6: format2)

* preamble: random access preamble

* ta: first timing advance

* nb_TB: number of enabled transport blocks

* MCSi_1: MCS index of the first transport block (-2: TB is disabled)

* MCSi_2: MCS index of the second transport block (-2: TB is disabled)

* TBS_1: TBS of first transport block (-2: TB is disabled)

* TBS_2: TBS of second transport block (-2: TB is disabled)

* nb_PRB: number of allocated PRBs

* connection_id: unique connection identifier

Dataset: DeepSlice & Secure5G - 5G & LTE Wireless Dataset

Published: November 2, 2019

Source (GitHub): https://github.com/adtmv7/DeepSlice

Source (Kaggle): https://www.kaggle.com/datasets/anuragthantharate/deepsliceDescription: Dataset that describes a Deep Learning model for 5G and Network Slicing (eMBB, URLLC, IoT)

supporting two different publications:

* DeepSlice: A Deep Learning Approach towards an Efficient and Reliable Network Slicing in 5G Networks

* Secure5G: A Deep Learning Framework Towards a Secure Network Slicing in 5G and Beyond

3.2.2 Dataset2: MILANO dataset

Source (Kaggle): https://www.kaggle.com/datasets/marcodena/mobile-phone-activity.

Description: The Mobile phone activity dataset is a part of the Telecom Italia Big Data Challenge, which is a rich and open multi-source aggregation of telecommunications, weather, news, social networks and electricity data from the city of Milan and the Province of Trentino (Italy). Every time a user engages a telecommunication interaction, a Radio Base Station (RBS) is assigned by the operator and delivers the communication through the network. Then, a new CDR is created recording the time of the interaction and the RBS which handled it. The following activities are present in the dataset:

* received SMS

* sent SMS

* incoming calls

* outgoing calls

* Internet activity

In particular, Internet activity is generated each time a user starts an Internet connection or ends an Internet connection. Moreover, during the same connection a CDR is generated if the connection lasts for more than 15 min or the user transferred more than 5 MB.

The datasets is spatially aggregated in a square cells grid. The area of Milan is composed of a grid overlay of 1,000 (squares with size of about 235×235 meters. This grid is projected with the WGS84 (EPSG:4326) standard. For more details we link the original paper http://go.nature.com/2fcOX5E.

3.3 PROJECT-SPECIFIC GENERATED DATA

3.3.1   RRU & AAA AW2S Products

Description: In the form of software and firmware packaged for the implementation of the planned proofs of concept. It will be towards the support of C/C++/VHDL applications

Standards: The developed code and corresponding data generated will follow 3GPP/IEEE/ISO standards. Moreover, all design will be developed with industry-standard tools for code development, compilation and execution

Data owner: AW2S and shared with the other parterns of the consortium according to the need of the project. Data provided via digital means (email, sharepoint) and formatted in word or excel.

Documentation/Metadata: not provided.

3.3.2   RAN stack software

Description: the developed software and there generated data related to the implementation of the RAN stack, with support to C/C++ applications. The development will be based on the existing srsRAN Project software.

Standards: The software is compatible with industry-standard C/C++ tools for code development, compilation and execution.

Data owner: Software Radio Systems, Ltd

License/data sharing. The software is made available through an open source, OSI-approved licence: AGPL-v3. AGPL-v3 is a strongly copy-left licence that places obligations on users who modify the codebase. The software developed in this project will be made available to the consortium partners through an online github repository. Where software features from this project are mainlined they will be made publicly accessible through github.

Access: Mainlined features are accessible on github and are publicly documented through on our srsRAN Project website: https://docs.srsran.com/en/latest/. SRS maintains repositories of its code here: https://github.com/srsran

Documentation/metadata: documentation provided to support the extension of the available modules and the related integration; no metadata will be provided.

3.4 DISSEMINATION AND PROMOTION DATASETS

The datasets generated for dissemination and promotion purposes are related to the <u>website</u>, <u>newsletter</u> and overall <u>event management</u>, whose description is detailed in the next subsections.

<u>Data collection</u>: In general, data collected are personal data, and in compliance with GDPR can only be used for the original scope they have been collected for: informing contacts about 5G-STARDUST. As such they will not be shared outside the parties in the consortium in charge of providing the newsletter services (so not to any external third party). It is also possible to unsubscribe from the newsletter at any time by contacting MARTEL/5G-STARDUST's consortium directly via the Contact section of the website, or through the website's GDPR Requests page: https://www.5g-stardust.eu/privacy-policy/gdpr-request/

<u>Data access</u>: On the other hand, the access to the website log data occurs via username and password. Only authorised users have the access and only admins at MARTEL can grant access. All personal data collected will be anonymised upon completion of the project and will be kept for no longer than two years after the completion of the project. Once the project is completed and after all reporting has been sent and approved by the EC, the consortium partners may decide to continue running any 5G-STARDUST active community platform, to transform it or to close it, but in any case, users will be asked about the re-use of their data. Access to data is possible only via authentication of previously authorised users by the data owner entity (MARTEL). Authentication and access channels to the data are encrypted as well.

<u>Data owner</u>: The single user retains the right and ownership to their personal data. MARTEL shall process and store the personal data of the data subject only for the period necessary to achieve the purpose of storage, or as far as this is granted by the European legislator or other legislators in laws or regulations to which MARTEL is subject to.

<u>Data curation and maintenance</u>: If the storage purpose is not applicable, or if a storage period prescribed by the European legislator or another competent legislator expires, the personal data are routinely blocked or erased in accordance with legal requirements. The criteria used to determine the period of storage of personal data is the respective statutory retention period. After expiration of that period, the corresponding data is routinely deleted.

3.4.1 5G-STARDUST website

The 5G-STARDUST website collects the following data:

- User account data: i.e. account of users authorised to publish content on the website. This are in general users part of MARTEL. This may include:

- Name and relevant titles

- Email address

- Contact data: i.e. data provided by website visitors filling in contact forms or sending email to "info@5g-stardust.eu" to request information to the 5G-STARDUST consortium. This may include:

    o   Name and relevant titles

    o   Email

- o Job title

- o Company name

- o Company address (in very rare cases)

- o Access log data: i.e. data that are collected by servers to control access to the web pages of the web site that may be used in the event of attacks and other cases requested by the law. This may include (1) the browser types and versions used, (2) the operating system used by the accessing system, (3) the website from which an accessing system reaches our website (so-called referrers), (4) the sub-websites, (5) the date and time of access to the Internet site, (6) an Internet Protocol address (IP address), (7) the Internet Service Provider of the accessing system, and (8) any other similar data and information that may be used in the event of attacks on our information technology systems.

Data is stored via WordPress (WP). WP has been used to build 5G-STARDUST's website. This content management system (CMS) uses the latest technology about PHP and MariaDB for the business logic and database respectively. WP provides al lot of plug ins in order to grant a great security both for the content and users. In fact, plugins such as anti-spam, anti-SQL injection, anti-brute force attack etc. can help to prevent spam and the most common attacks. Moreover, WP provides different access roles, in order to grant the right permissions to the right users. Users' WP passwords are encrypted through RSA technology, so no one can decrypt them. Neither a WP administrator.

Users can access the website only via an encrypted connection (https), in order to add a second security layer between the user and the website itself. Data is stored on WordPress.

3.4.2 5G-STARDUST Newsletter

We use the third-party service provider Mailerlite to process and store the data which 5G-STARDUST uses to manage the newsletter subscriber lists and send emails to our subscribers. We don't share users data with any other third-party.

Types and formats of data collected:

- E-mail address of visitors registering to the newsletter.

- During the registration for the newsletter, we also store the IP address of the computer system assigned by the Internet Service Provider (ISP) and used by the data subject at the time of the registration, as well as the date and time of the registration.

- The newsletter of 5G-STARDUST contains so-called tracking pixels. A tracking pixel is a miniature graphic embedded in such e-mails, which are sent in HTML format, to enable log file recording and analysis. Based on the embedded tracking pixel, MARTEL may see if and when an e-mail was opened by a data subject, and which links in the e-mail were called up by data subjects.

Data are stored in a GDPR compliant newsletter service (Mailerlite) and are encrypted.

3.4.3 5G-STARDUST Events

For users registering to events we may collect:

- Name and relevant titles

- Email

- Job title

- Company name

- Company address (in very rare cases)

Data and metadata will be requested, stored, and transferred in comma-separated values CSV format.

Event Participation: We use third-party service provider Ti.to to process and store data. We don't share user's data with any other third-party.

3.5 SUMMARY

Based on the preliminary survey and the related data collected through the questionnaire available in Appendix A, it is possible to summarise the current status of data planned to be used as of June 30th 2023, which constitutes the baseline for the current version of the DMP. As formerly mentioned, the DMP will be periodically updated so as to reflect the evolutions of the project with respect to data generation and usage, which are expected to happen as soon as the technical development workpackages (WP4 and WP5) will start, i.e. in July and August 2023 respectively.

Table 1 Datasets Summary

Data Nr.	Responsible Beneficiary	Data set reference and name and [used methodology]	Data set description			Research data identification				
			End user (e.g. university, research organization, SME's, scientific publication)	Existence of similar data (link, information)	Possibility for integration and reuse (Y/N) + information	D [2]	A [3]	AI [4]	U [5]	I [6]
1	CTTC	ADOPNET	University	-	Y	x	x	x	-	x
2	CTTC	MILANO	University	-	Y	x	x	x	-	X
3	MAR	Website	SME	-	N	x	x	x	-	x

---

[2] Discoverable

[3] Accessible

[4] Assessable and intelligible

[5] Usable beyond the original purpose of which it was collected

[6] Interoperable to specific quality standards

Dat a Nr.	Responsibl e Beneficiar y	Data set reference and name and [used methodolog y]	Data set description			Research data identification				
			End user (e.g. university, research organizatio n, SME's, scientific publication)	Existence of similar data (link, informatio n)	Possibility for integratio n and reuse (Y/N) + informatio n	D 2	A 3	AI 4	U 5	I 6
4	MAR	Newletter	SME	-	N	x	x	x	-	x
5	MAR	Event managemen t	SME	-	N	x	x	x	-	X
6	AW2S	RRU&AAA	SME	-	N	-	x	x	-	x
7	SRS	RAN stack	SME	-	Y	x	x	x	-	x

As highlighted in the previous sections, the current summary does not consider the datasets that will be made available by ORA and HSP in the course of the project, whose definition is not yet finalised nor is the data processing needed for the actual use in the corresponding work packages 4 and 5. These will be directly reported in the D4.1 [1] and D5.1 [2], expected to be issued by the end of the year.

Another important note we can draw from the above table is that the list of data generated by the project is limited to 2 public datasets and those possibly generated by the products owned by AW2S and SRS. It is expected that this list will become more consolidated (i.e. grow in number) as soon as the development phase will start and therefore clearer will be which products or modules will be developed and integrated together. Likewise, as soon as simulation data will be available, this will represent an additional point to be reported in future evolutions of the DMP.

4 DATA MANAGEMENT IN H2020/HE PROGRAMMES

According to the European Commission (EC) all project proposals submitted to "Research and Innovation actions", "Innovation actions" and "Coordination support actions" have to include a section on research data management which is evaluated under the criterion 'Impact'. Projects participating in the pilot action on open access to research data have to develop a DMP) to specify what data will be open[7].1

The DMP is defined as:

> *"Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. The use of a Data Management Plan is required for projects participating in the Open Research Data Pilot. Other projects are invited to submit a Data Management Plan if relevant for their planned research."*

The purpose of a DMP is to provide a discussion of the main elements of the data management policy that will be used by the applicants with regard to all the datasets that will be generated by the project.

Table 2 Clarification of terms

Data classification	Definition
Research data	Research data is the evidence that underpins all research conclusions (except those which are purely theoretical) and includes data that have been collected, observed, generated, created or obtained from commercial, government or other sources, for subsequent analysis and synthesis to produce original research results. These results are then used to produce research papers and submitted for publication.
Open research data	Openly accessible research data can typically be accessed, mined, exploited, reproduced and disseminated, free of charge for the user.
Secondary data	Secondary data are data that already exist, regardless of the research to be conducted.
Open access	Open access is understood as the principle that research data should be accessible to relevant users, on equal terms, and at the lowest possible cost. Access should be easy, user-friendly and, if possible, Internet-based.

---

7   http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Metadata	Metadata is data used to describe other data. It summarizes basic information about data, which can make finding and working with instances of data easier.
Research data repositories	Research data repositories are online archives for research data. They can be subject based/thematic, institutional or centralized

Overall, having considered all relevant principles regarding lawful processing of personal data, scientific research data should be easily discoverable, accessible, assessable and intelligible, useable beyond the original purpose for which it was collected and interoperable to specific quality standards.

The 5G-STARDUST Data Management also follows the Guidelines on FAIR (Findability, Accessibility, Interoperability, and Reuse) Data Management in Horizon 2020 and further reinforced under the Horizon Europe programme, released by the European Commission Directorate – General for Research & Innovation. This Horizon 2020 FAIR DMP template[8] has been designed to be applicable to any Horizon 2020 project that produces, collects or processes research data. Later an additional template has been issued to cover projects funded in the HE programme[9], essentially aligned with the previous template. According to these guidelines the management and organization of data should be based on four basic principles, which determine how research outputs should be processed so that they can be more easily accessed, understood, exchanged and reused. This means that data must be findable, accessible, interoperable and re-useable, for example by researchers interested in using the data in further research in the field.

The aim of these principles is to make digital assets more easily discoverable, accessible, interoperable, and reusable for researchers and practitioners. By following the FAIR principles, data becomes more valuable and usable, enabling better collaboration and knowledge sharing. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation-solution. EC provides a Template with the FAIR principle. This template is not intended as a strict technical implementation of the FAIR principles, it is rather inspired by FAIR as a general concept. The template represents the set of questions that someone should answer with a level of detail appropriate to the project.

In more detail, the considered measures complying to FAIR data management include:

- **Findability:** It includes the use of standards and metadata format, persistent and unique identifiers, naming conventions, search keywords, and version numbers, which help to make data discoverable and accessible. Findability ensures that data can be easily located and retrieved, allowing others to easily access, reuse, and build upon it.
- **Accessibility:** It refers to the ease with which data can be located and retrieved by both humans and machines. It involves making data available and usable to anyone who needs it, without any barriers such as technical, legal, or institutional restrictions.

---

[8] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 Version 3.0, 26 July 2016, available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[9] HE Data Management template: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/temp-form/report/data-management-plan_he_en.docx

This includes making data available in a machine-readable format and providing adequate documentation of the data. Accessibility also requires data to be properly stored and maintained, so that it remains usable over time. By ensuring accessibility, FAIR data management helps promote data reuse and scientific collaboration.

- **Interoperability** refers to the ability of different systems, platforms, and tools to work together and exchange data effectively. This means that data should be stored, formatted, and described in such a way that it can be understood and used by different systems and applications, regardless of the technology used to create or store the data. Interoperability is an important aspect of FAIR data management because it enables seamless data exchange between different systems and facilitates data reuse, which is crucial for scientific collaboration and discovery. Interoperability helps to ensure that data is accessible, usable, and valuable over time, and it supports the efficient integration of new data and tools into existing systems.

- **Reusability**: Reusability in FAIR data management refers to the ability to use data in multiple ways, for multiple purposes, with proper attribution and consent, while ensuring that the data remains secure and respects the privacy of individuals. This principle is an important aspect of FAIR data management as it helps to maximise the value of data, promote scientific progress, and support innovation while still respecting ethical considerations such as privacy and confidentiality.

- **Resource allocation:** costs of making data FAIR, how to cover these costs, the role of the data manager, partner responsibilities, the potential value of long-term preservation, and preservation costs.

- **Archiving and preservation**: at the end of the project, data selection for archiving and preservation, estimated final volume, recommended preservation duration, and long-term preservation storage.

- **Data security:** data security provisions and security of long-term preservation.

- **Ethical considerations:** Impact of ethical or legal issues. To ensure a user-friendly approach to the preparation and maintenance of the data.

5 DATA PUBLICATION

5.1 OVERVIEW

The approach adopted in the project with respect to data publication is twofold and consists in

1. developing a Data Management Plan (DMP), as deemed obligatory by EC for the projects funded under the Horizon Europe programme, to which 5G-STARDUST belongs

2. providing open access to research data, as far as possible

In this respect, the project aims at

- Developing (and keeping up-to-date) DMP through the evolution that the data management process may be subject to, upon emergence of new project needs or availability/use of data.

- Deposit the data in a research data repository.

- Ensure third parties can freely access, mine, exploit, reproduce and disseminate this data.

- Provide related information and identify (or provide) the tools needed to use the raw data to validate the research.

This pragmatic framework will be applied to:

- The data (and metadata) needed to validate results in scientific publications.

- Other curated and/or raw data (and metadata) that are specified in the DMP.

Might be excluded from this methodology all projects artefacts for which specific beneficiary policies or limitation on the publication may arise.

5.2 PUBLISHING INFRASTRUCTURE FOR OPEN ACCESS

5.2.1 General Approach

The 5G-STARDUST publication infrastructure consists of a process and a few web-based publication platforms that together provide long-term open access to all publishable, generated or collected results of the project. The implementation of the project will be done in accordance with the applicable regulations in national and EU level and, especially, with the General Data Protection Regulation (GDPR)[10] protection of personal data.

---

[10] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

In more detail, the project is not envisaging any specific cases (i.e. data generated in the project by the developed modules or data provided by the consortium or third parties) where personal data information or sensitive information of internet users is collected (IP addresses, email addresses or other personal information) or further elaborated for the objectives of the project. On the contrary, where applicable, anonymisation procedure will be applied.

In the case, the project during its forthcoming development see the need for collecting or re-elaborating personal data, this will be performed in compliancy to the GPDR regulations. In such a case, it is affirmed that use and re-elaboration of necessary personal data will be towards the achievement of some planned project objectives.

The overall publishing process and the implied used web-based platforms as well as other tolls (as far as applicable) are described in the following subsections on the basis

5.2.2 Publishing Process

The publishing process followed by 5G-STARDUST is based on an usual framework related to the innovation, quality, and soundness of the proposed publication activity. Moreover, specific regulations as to the timing, the approval, and check for possible conflict in terms of ethics or IPRs with other partners is directly addressed in the project consortium agreement (CA). In general, the publication process is applied to any possible artefacts generated during the project, such as white papers, brochure, leaflets, presentations, scientific publications, and anonymous usage data. By following the process shortly outlined in the following, each result is either classified public or non-public. Public means that the result can be published under the open access policy, whereas non-public means that it must not be published.

In more detail, the publication process is regulated by a few questions, whose answer are used to define the corresponding classification of a potential publication. Such a process is run by the project coordinator at the time approval for a publication is sought, in cooperation with the project executive board and the publication authors, where considered necessary or helpful for completing the validation process in a fair and effective manner.

The following questions must be answered towards a publication classification:

Q1: Does a result provide significant value to others or convey a relevant scientific conclusion to a given subject or problem?

YES: it is classified as public.

NO: it is classified as non-public, because it's about scientific finding of limited interest or no possible reuse/applicability from the community, being very specific to the project implementation

Q2: Does a result include personal information that is not the author's name?

YES: It is classified as non-public. Personal information besides the name of an author (co-author, or acknowledged person who has supported the preparation of this publication) must be removed for a publication. As such, removal privacy-related information is a necessary condition for making a 5G-STARDUST result public.

NO: it is classified as public.

Q3: Does a result allow the identification of individuals even without the name?

YES: the result is classified as non-public. Sometimes data inference allows to indirectly reveal the identity of user. As such, publication gran be granted only if the included information does not allow single individuals to be identified. To this end, anonymisation procedures can be carried out.

NO: the result is classified as public.

Q4: Does a result name technology that is part of an ongoing, project-related patent application?

YES: result is classified as non-public. Results will be be published after patent has been filed, if this is still considered meaningful (because of patent filing process duration) and desirable.

NO: the result is classified as public.

Q5: Can a result upon publication result in abuse, misuse, or infringement of societal norms and overall ethics?

YES: the result is classified as non-public. Moreover, the project coordinator will be urged to check the reason for such a criticality, in discussion with the executive board and the result authors.

NO: the result is classified as public.

Q6: Does a result include business or trade secrets of one or more partners?

YES: the result is classified as non-public, unless the involved/affected partners are not against such a publication. Additional regulations from the CA may also apply here. Differently, requirement for the publication will be that business or trade secrets are removed according to needs expressed by the project consortium.

NO: the result is classified as public.

Q7: Does a result break national security interests for any project partner?

YES: the result is classified as non-public.

NO: the result is classified as public

As a conclusion of the process, if all questions answers yield a "public" classification, then the clearance process is successful and a given result can be published. If at least one answer result in "non-public", then the process is considered unsuccessful and the decision is notified

to the authors to proceed accordingly or to revise accordingly (if doable) the paper (for example) in order to meet the aforementioned publication criteria.

5.2.3 Publishing Platform

The 5G-STARDUST project features a few platforms to publish results openly. As follows the tentative list (i.e. as available or planned until end of June 2023) of platforms planned for use during the project, for which the specific characteristics with respect to publishing, storage, and backup capabilities are also given.

5.2.3.1 Project Website

The 5G-STARDUST project has been operation since the first weeks of the project, in order to promote in first instance the objectives and scope of the project, and then later to offer a space where to report the latest achievements and then provide interested people with published material. As such, the website in structured in a few sections dealing with the overall project characteristics and key achievements, activity updates and overall news populated on a regular basis, subscription to newsletters, and then a dedicated area for downloads of public papers (scientific papers, presentations, brochures, reports) and deliverables. More importantly, scientific publications are supposed to appear either in pre-camera ready form where possible or through links to the publisher's websites in case these are not open access. All documents are published using the portable document format (PDF). All downloads are enriched by using simple metadata information, such as the title, type of the document, publication time, etc. The website is hosted by partner MARTEL. All webpage-related data is backed up on a regular basis. All information on the project website can be accessed without creating an account. The website is backed up once per month.

5.2.3.2 Social media and other media containers

Social media are being regularly used (i.e. Linkedin™, Twitter™) for reporting the appearance of the project in key events and collecting interest from the relevance audience. As such, posts may link to public documents developed and made available from the project consortium.

A YouTube™ channel is also being set up to store presentation and talks given in key events and hence raise the awareness of the interested audience.

5.2.3.3 Project Sharepoint

The project features a dedicated sharepoint for regular operations of the project (i.e. data collection and project deliverable preparation). The sharepoint is regulated by private access (i.e. through credentials), although some document there stored could be made available by making specific files accessible from parties external to the project consortium by means of dedicated links.

5.2.3.4 GitLab

GitLab is a well-established online repository for distributed source code development, management, sharing, and revision control. It is primarily used for source code data. It enables world-wide collaboration between developers and provides also some facilities to work on documentation and to track issues. GitLab provides paid and free service plans. Free service plans can have any number of public, open-access repositories with unlimited collaborators. Private, non-public repositories require a paid service plan. Many open-source projects use GitLab to share their results for free. The platform uses metadata like contributors' nicknames, keywords, time, and data file types to structure the projects and their results. The terms of

service state that no intellectual property rights are claimed by GitLab over provided material. For textual metadata items, English is preferred.

Where applicable (depending on the beneficiary policy), source-code components that are implemented during this project and decided to be public will be uploaded to an open access GitLab repository.

5.3 DATA ACCESS AND SHARING

The accessing and sharing of data is firstly ruled by two documents: The Consortium Agreement (CA), which stipulates under which conditions transmitted information between the project partners is deemed confidential and must not be further disseminated; and the Description of Action (DoA) which stipulates the dissemination level of each deliverable. Moreover, the project consortium will comply with the FAIR (findable, accessible, interoperable and reusable) guidelines of the H2020 programme and then further reinforced as part of Horizon Europe programme.

The data necessary to successfully fulfill the objectives of the projects across the project Work Packages (WPs) are without any restrictions amongst the WP and tasks partners either via internal repositories (project sharepoint) or direct communication (i.e. via emails or other digital means). Public data will be made available at the project's website or other repositories, as appropriate. Users will be made aware of this data primarily through research publications, patent applications, dissemination activities, invited talks, social networks and the project website. Data will be made available to the project consortium as soon as available through dedicated announcements happening through the proper forms; to standardization bodies when required (i.e. for contribution to given standardisation actions); and to the public at the due date of the derivable, and, in case a research publication is based on that, as soon as the paper is published. If access to confidential data is necessary by the public, restrictive measures will be put in place, by regulating the access through credentials provided to the interested users (where possible and agreed between the involved parties).

Each of the previously defined (data and/or research) artefacts has its own set of questions that have to be addressed. The proposed template (appearing also at the end of this document in a dedicated Appendix) states that no detailed answers are expected for all the questions for the present version of the DMP. On the contrary, the DMP is intended to be a living document hence subject to extensions and updates on the basis of the project implementation progress and on the usage and generation of data as the case emerge in the course of the project. Moreover, as formerly reported, dataset being used in the projects will be object of dedicated deliverables and hence neither reported in detailed manner in updated version of the DMP.

6 ETHICS

At the time of the grant agreement preparation, no specific ethics' related issues were raised by EC, so that no need for a dedicated ethics board nor for adding dedicated deliverables was appearing. Nevertheless, the project coordinator with the support of the project executive board is in charge of assessing any risk related to the publication of data that may result in some ethics criticality.

7   CONCLUSIONS

This document reflects the vision of the 5G-STARDUST consortium for the generation and handling data as of July 1st, 2023. It is foreseen that most of the generated data will be related to source-code and the data necessary to support the demonstrators, the documentation, and the experimental results, as well as data resulting from the needed simulation campaigns to validate the proposed technology concepts.

Efforts will be put in place, such as the adherence of the developed code as much as possible with consolidated standard and current best practice, so that the 5G-STARDUST mechanisms can be generic enough to be able to execute on a wide gamut of heterogeneous devices. In addition, other generated datasets, like experimental results, will follow standardised formats, and will be documented in forthcoming deliverables as part of WPs 4 and 5, so that they can be easily used as reference in other research projects. The datasets will be disseminated through dedicated deliverables later in the project, research publications, patent applications, invited talks, among others, and will be preserved at the partners repositories for at least 3 years.

It should be noted that since this is a preliminary version of the DMP and no data have been generated nor used so far apart those for promotional and dissemination purposes, this document presents preliminary proposals in terms of sharing, data management, and archiving. The DMP will be updated periodically during the project to reflect changes in the properties of the data that may be made available by the project, and to add more concrete information about the datasets. Such updates will be published on the project website, as no additional deliverable on this matter is envisioned in the workplan of the project.

REFERENCES

[1] D4.1 – Open data sets for ML-based RRM

[2] D5.1 – Open Data Sets for AI Data Driven Networking

APPENDIX A – DMP QUESTIONNAIRE

7.1 DATASETS

- Do you plan to generate/gather data (measurements, algorithms, source code, software stacks, APIs, policies, etc.) inside the project or during the project lifetime?
  <fill in>

- If yes, please specify what type, format and amount of data.
  <fill in>

- Do you use any existing data/datasets, which you can/will reuse for further developments/implementations inside the project or during the project lifetime?
  <fill in>

- Do you follow any standards for such data? (NIST, ISO, etc.)
  <fill in>

- What information is needed for the data to be read and interpreted in the future?
  <fill in>

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.

*********************Please FILL BELOW *************************

7.2 INTELLECTUAL PROPERTY RIGHTS

- Who owns the data?
  <fill in>

- Will the data be licensed for reuse? If yes, under which licence?
  <fill in>

- Are there any restrictions on the reuse of existing/third-party data?
  <fill in>

- Will data sharing be postponed / restricted e.g. to publish or seek patents?

7.3 ACCESS AND SHARING

- With whom will you share the data (e.g. consortium only, publicly accessible, etc.), and under what conditions/restrictions?
  <fill in>

- How will access to the data be provided in case of restrictions?
  <fill in>

- Are there any ethical or legal issues that can have an impact on data sharing? If yes, describe them and the way they are dealt with.
  <fill in>

- How will potential users find out about your data?
  <fill in>

- Will you share data via a repository, handle requests directly or use another mechanism?
  <fill in>

- When will you make the data available?
  <fill in>

- Will you pursue getting a persistent identifier for your data?
  <fill in>

- Do your chosen types, formats and software enable sharing and long-term access of the data?
  <fill in>

7.4 ARCHIVING AND PRESERVATION

- What data must be retained/destroyed for contractual, legal, or regulatory purposes?
  <fill in>

- How will you decide what other data to keep?
  <fill in>

- What are the foreseeable research uses for the data?
  <fill in>

- How long will the data be retained and preserved?
  <fill in>

- Where e.g. in which repository or archive will the data be held?
  <fill in>

- What costs if any will your selected data repository or archive charge?
  <fill in>

- Have you costed in time and effort to prepare the data for sharing / preservation?
  <fill in>

7.5 DOCUMENTATION & METADATA

- Do you plan to create/provide documentation?
  <fill in>

- Do you plan to create/provide metadata?
  <fill in>

- If so, how will you capture / create this documentation and metadata?
  <fill in>

- What metadata standards will you use and why?
  <fill in>

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 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.

*********************Please FILL BELOW ************************