



European
Commission

Horizon 2020
European Union funding
for Research & Innovation

Big Data technologies and extreme-scale analytics



Multimodal Extreme Scale Data Analytics for Smart Cities Environments

D2.5: Corpus-as-a-Service specifications and business continuity[†]

Abstract: This deliverable is the main report for the MARVEL Data Corpus and the related activities that were performed. It provides details concerning: *i*) the design and implementation of the service, *ii*) the deployment in the cloud infrastructure of PSNC, *iii*) the security and privacy assessments, *iv*) licencing aspects, *v*) a summary of the legal compliance (detailed in D2.6), *vi*) user evaluation trials and acceptance, *vii*) community building and Corpus continuity plans after the end of the project, and *viii*) an outline of the exploitation strategy (detailed in D7.6). D2.5 is the final outcome of the tasks T2.3 (M30) and T2.4 (M36).

Contractual Date of Delivery	31/12/2023
Actual Date of Delivery	04/01/2024
Deliverable Security Class	Public
Editors	<i>Kostas Poulios (STS)</i> <i>George Hatzivasilis (FORTH)</i>
Contributors	IFAG, AU, ATOS, CNR, FBK, AUD, MT, UNS, ITML, GRN, ZELUS
Quality Assurance	<i>Tomas Pariente Lobo (ATOS)</i> <i>Emmanouil Michalodimitrakis (FORTH)</i>

[†] The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337.

The *MARVEL* Consortium

Part. No.	Participant organisation name	Participant Short Name	Role	Country
1	FOUNDATION FOR RESEARCH AND TECHNOLOGY HELLAS	FORTH	Coordinator	EL
2	INFINEON TECHNOLOGIES AG	IFAG	Principal Contractor	DE
3	AARHUS UNIVERSITET	AU	Principal Contractor	DK
4	ATOS SPAIN SA	ATOS	Principal Contractor	ES
5	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR	Principal Contractor	IT
6	INTRASOFT INTERNATIONAL S.A.	INTRA	Principal Contractor	LU
7	FONDAZIONE BRUNO KESSLER	FBK	Principal Contractor	IT
8	AUDEERING GMBH	AUD	Principal Contractor	DE
9	TAMPERE UNIVERSITY	TAU	Principal Contractor	FI
10	PRIVANOVA SAS	PN	Principal Contractor	FR
11	SPHYNX TECHNOLOGY SOLUTIONS AG	STS	Principal Contractor	CH
12	COMUNE DI TRENTO	MT	Principal Contractor	IT
13	UNIVERZITET U NOVOM SADU FAKULTET TEHNICKIH NAUKA	UNS	Principal Contractor	RS
14	INFORMATION TECHNOLOGY FOR MARKET LEADERSHIP	ITML	Principal Contractor	EL
15	GREENROADS LIMITED	GRN	Principal Contractor	MT
16	ZELUS IKE	ZELUS	Principal Contractor	EL
17	INSTYTUT CHEMII BIOORGANICZNEJ POLSKIEJ AKADEMII NAUK	PSNC	Principal Contractor	PL

Document Revisions & Quality Assurance

Internal Reviewers

1. *Tomas Pariente Lobo, ATOS*
2. *Emmanouil Michalodimitrakis, FORTH*

Revisions

Version	Date	By	Overview
4.0	04/01/2024	STS	Final version for submission
3.1	31/12/2023	FORTH	Comments on the third draft
3.0	28/12/2023	STS	The third draft
2.4	28/12/2023	FORTH	Comments on the second draft
2.3	27/12/2023	STS	The second draft
2.2	26/12/2023	ATOS	Comments of the first draft
2.1	23/12/2023	STS	The first draft
2.0	01/12/2023	STS	Final ToC
1.1	25/10/2023	STPM	Comments on the ToC
1.0	16/10/2023	STS	ToC

Disclaimer

The work described in this document has been conducted within the MARVEL project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957337. This document does not reflect the opinion of the European Union, and the European Union is not responsible for any use that might be made of the information contained therein.

This document contains information that is proprietary to the MARVEL Consortium partners. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to any third party, in whole or in parts, except with prior written consent of the MARVEL Consortium.

Table of Contents

LIST OF FIGURES.....	6
LIST OF TABLES.....	7
LIST OF ABBREVIATIONS.....	8
EXECUTIVE SUMMARY	10
1 INTRODUCTION.....	12
1.1 PURPOSE AND SCOPE.....	15
1.2 RELATION TO OTHER WORK PACKAGES, DELIVERABLES, AND ACTIVITIES.....	15
1.3 STRUCTURE OF THE DELIVERABLE.....	16
2 OVERVIEW OF THE MARVEL DATA CORPUS	17
2.1 GOALS, OBJECTIVES, AND VISION	17
2.2 CURRENT STATUS AND AVAILABLE DATA	19
2.3 MAIN TECHNOLOGIES USED	19
2.3.1 <i>Hadoop repository</i>	20
2.3.2 <i>HBase database and HDFS</i>	20
2.3.3 <i>Angular 2.0 for GUI development</i>	20
2.3.4 <i>Augmentations</i>	21
2.3.5 <i>Anonymisation</i>	21
2.3.6 <i>AI inference and Data management</i>	22
2.4 IMPLEMENTATION AND EVALUATION STAGES.....	23
2.4.1 <i>Implementation roadmap</i>	23
2.4.2 <i>Performance and benchmarking</i>	23
2.4.3 <i>User acceptance</i>	23
2.4.4 <i>Security assessment</i>	24
2.4.5 <i>Privacy assessment</i>	24
2.5 LEGAL ASPECTS.....	24
2.6 EXPLOITATION STRATEGY	25
2.7 LESSONS LEARNED	26
3 TECHNICAL DESCRIPTION OF THE MARVEL DATA CORPUS DATABASE	28
3.1 INFRASTRUCTURE & DEPLOYMENT	28
3.2 IMPLEMENTATION OF DATABASE APPLICATION AND SERVICES	31
3.3 INTEGRATION WITH THE MARVEL PLATFORM.....	33
3.3.1 <i>Ingestion of anonymised piloting data from Edge-Fog-Cloud</i>	34
3.3.2 <i>Ingestion of real-time data with identified events from the AI inference pipeline via StreamHandler</i>	34
3.3.3 <i>Ingestion of augmented data</i>	36
3.3.4 <i>Ingestion of data from the MARVEL users' Web Interface</i>	36
3.4 MAIN SECURITY AND PRIVACY ASPECTS.....	37
4 USER INTERFACES	38
4.1 PUBLIC WEB INTERFACE.....	38
4.1.1 <i>Introductory Page</i>	38
4.2 DATASETS VIEW.....	39
4.3 DATASET DETAILED VIEW.....	40
4.4 SNIPPETS VIEW.....	42
4.5 SNIPPET DETAILED VIEW	42
4.6 MARVEL USER WEB INTERFACE.....	42
4.7 PROGRAMMATIC INTERFACES & APIS	43
4.7.1 <i>Public APIs</i>	43
4.7.1.1 <i>Public APIs for Datasets</i>	43
4.7.1.2 <i>Public APIs for Snippets</i>	44
4.7.2 <i>Private APIs</i>	44

5	AUGMENTATION ENGINE	45
5.1	AUGMENTATIONS IN ML	45
5.2	IMPLEMENTATION AND SUPPORTED AUGMENTERS	45
6	THE INGESTED DATASETS	50
6.1	USE CASES	50
6.2	STREAM SOURCES	50
6.3	DATASETS	53
7	EVALUATION STUDIES	55
7.1	PERFORMANCE/BENCHMARKING	55
7.2	EASE-OF-USE & USER SATISFACTION	55
7.2.1	<i>Evaluation Study 1 – Internal</i>	55
7.2.2	<i>Evaluation Study 2 – Internal</i>	56
7.2.3	<i>Evaluation Study 3 – External</i>	56
7.3	SECURITY ASSESSMENTS	56
7.3.1	<i>Assurance Platform description</i>	56
7.3.1.1	Core Platform	57
7.3.1.2	Asset-based vulnerability assessment	57
7.3.1.3	Dynamic Tester	57
7.3.1.4	EEvent REaSonng Toolkit (EVEREST)	58
7.3.1.5	Event Captors	58
7.3.1.6	Security Component	58
7.3.2	<i>Vulnerability assessments</i>	58
7.3.3	<i>Assurance Profiles and CIA Monitoring</i>	60
7.4	PRIVACY ASSESSMENTS	61
7.4.1	<i>SENTINEL Platform description</i>	61
7.4.1.1	SENTINEL Platform methodology and components	62
7.4.1.2	Privacy assessment	63
7.4.2	<i>Ethics, privacy, and data protection compliance monitoring</i>	63
8	COMMUNITY BUILDING AND ACTIONS TOWARDS THE DATA ECONOMY VISION OF SMART CITIES	65
8.1	UNIVERSITIES AND COLLEGES	65
8.2	RESEARCH INSTITUTES	66
8.3	SMES	67
8.4	INDUSTRY	68
9	CONCLUSION	70
10	REFERENCES	71
11	ANNEX 1 – DATA MODEL	74
12	ANNEX 2 – LETTER OF INTENT (LOI) TEMPLATE	77
13	ANNEX 3 – USER SATISFACTION STUDIES	79
13.1	FIRST EVALUATION BY INTERNAL USERS	79
13.1.1	<i>Performance Evaluation section</i>	79
13.1.2	<i>Usability Evaluation section</i>	81
13.1.3	<i>User-friendliness Evaluation section</i>	83
13.1.4	<i>Use and Exploitation Evaluation section</i>	84
13.1.5	<i>Demographics section</i>	85
13.2	SECOND EVALUATION BY INTERNAL USERS	86
13.2.1	<i>Performance, Usability, and User-friendliness Evaluation section</i>	86
13.2.2	<i>Use and Exploitation section</i>	88
13.2.3	<i>Demographics section</i>	89
13.3	THIRD EVALUATION BY EXTERNAL USERS	90
13.3.1	<i>Trail and Evaluation section</i>	90
13.3.2	<i>Use and Exploitation section</i>	92
13.3.3	<i>Demographics section</i>	94
14	ANNEX 4 – JUSTIFICATION FOR THE TOTAL DATA CORPUS SIZE	96

List of Figures

Figure 1. MARVEL Data Corpus main concepts	18
Figure 2. The Data Corpus VMs status.....	28
Figure 3. Main building blocks of the core Data Corpus repository implementation.....	29
Figure 4. Infrastructure and Deployment view on several VMs.....	30
Figure 5. MARVEL conceptual architecture	33
Figure 6. Integration of Data Corpus with the StreamHandler and AI inference pipeline	34
Figure 7. StreamHandler’s output on folders.....	35
Figure 8. Create dataset from private GUI.....	36
Figure 9. GUI – Introductory Page 1-1.....	38
Figure 10. GUI – Introductory Page 1-2.....	39
Figure 11. GUI – Introductory Page 1-3.....	39
Figure 12. GUI – Datasets View 1-1 – Datasets’ Aggregated data and Table	40
Figure 13. GUI – Datasets View 1-2 – Complex Search tab	40
Figure 14. GUI – Dataset View 1-1 – Descriptive metadata	41
Figure 15. GUI – Dataset View 1-2 – Technical metadata.....	41
Figure 16. GUI – Dataset View 1-3 – Snippets’ list	41
Figure 17. GUI – Snippets View – Snippets’ Aggregated data and Table	42
Figure 18. GUI – Snippet View – Technical metadata and Download button	42
Figure 20. Original and Augmented datasets for one Gozo stream.....	49
Figure 21. Total snippets volume.....	54
Figure 22. Assurance Platform – Internal architecture	57
Figure 23. Assurance Platform GUI – Availability monitoring.....	61
Figure 24. SENTINEL Platform – Assessment scope according Privacy Risk level.....	62
Figure 25. SENTINEL Platform – Privacy assessment.....	63

List of Tables

Table 1. KPIs evaluation	18
Table 2. Configuration of the Data Corpus VMs	30
Table 3. API – View all Datasets	43
Table 4. API – Search Datasets	43
Table 5. API – View specific Dataset	43
Table 6. API – View all Snippets	44
Table 7. API – View specific Snippet	44
Table 8. API – Download specific Snippet	44
Table 9. Video and Audio Augmenters	46
Table 10. Stream sources	51
Table 11. Performance benchmarking	55
Table 12. Assurance Platform vulnerability assessment – Identified CVEs	59

DRAFT

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AV	Audio/Video
BDVA	Big Data Value Association
CCTV	Closed-Circuit Television
CIA	Confidentiality, Integrity, and Availability
CLI	Command Line Interface
CPE	Common Platform Enumeration
CSA	Compliance Self-Assessment
CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
CWE	Common Weakness Enumeration
dBFS	decibels relative to full scale
DFB	Data Fusion Bus
DMT	Decision-Making Toolkit
DPIA	Data Protection Impact Assessment
DPMAN	Data Protection Management
E2F2C	Edge-to-Fog-to-Cloud
EAB	Ethics Advisory Board
ELK	Elasticsearch, Logstash, and Kibana
EU	European Union
EVEREST	Event REasoning Toolkit
fps	Frames per second
GAN	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GUI	Graphical User Interface
HDFS	Hadoop Distributed File System
HPC	High-Performance Computing
IaaS	Infrastructure-as-a-Service
ICT	Information and Communications Technology
IDS	Intrusion Detection System
IT	Information Technology
JSON	JavaScript Object Notation
KER	Key Exploitable Result
KPI	Key Performance Indicator

LOI	Letter of Intent
LPC	Linear Predictive Coding
MQTT	Message Queuing Telemetry Transport
NoSQL	not only SQL or non-SQL
NVD	National Vulnerability Database
OTM	Operational and Technical Measure
PA	Processing Activity
PaaS	Platform-as-a-Service
PDF	Portable Document Format
PDLM	Personal Data Lifecycle Management
PIL	Python Imaging Library
REST	REpresentational State Transfer
RTSP	Real-Time Streaming Protocol
SBOM	Software Bill of Materials
SLA	Service Level Agreement
SME	Small and Medium-sized Enterprise
SNR	Signal-to-Noise Ratio
SOC	Security Operations Center
SSH	Secure shell
SSL	Secure Sockets Layer
SSO	Single Sign-On
TLS	Transport Layer Security
ToE	Target of Evaluation
TRL	Technology Readiness Level
TXT	TeXT file extension
UI	User Interface
VAD	Voice Activity Detection
VM	Virtual Machine
VPN	Virtual Private Network
WP	Work Package
XML	Extensible Markup Language

Executive Summary

This deliverable (*D2.5 – Corpus-as-a-Service specifications and business continuity*) details the progress concerning the MARVEL Data Corpus. This Corpus is a *Big Data repository that maintains the datasets that were selected during the course of the project* and are meant to be shared publicly with research and industrial communities. This document mainly provides information for the technical aspects, the datasets that have been structured, the security and privacy analysis on the deployed infrastructure, the user evaluations and the community of research and industrial partners that has been established, and the operational plan for up to one year after the end of the project. For completion, a summary of the legal compliance aspects is provided along with an outline of the exploitation plan, which are both detailed in other deliverables, also due at M36.

The MARVEL Data Corpus has been *deployed in the cloud infrastructure* provided by PSNC. PSNC aims to serve as a Polish Sound Data Hub that facilitates third parties to deploy applications based on environmental multimodal analytics. The repository has been *installed across several Virtual Machines (VMs)* acting as data nodes, with the total size *reaching the 1.IPBs* at M36. The Corpus has been *implemented as a service with around 35 Application Programmable Interfaces (APIs)*. Apart from the *programmable interface, two Graphical User Interfaces (GUIs)* were developed *for internal and external users*, respectively. *Input from around 29 streaming sources* has been integrated *from the three pilots* (MT, GRN, and UNS), as well as the *collaborating smart city of Gozo* in Malta. Around *52,000 snippet files* (audio, video, and audio-video) have been recorded within *64 datasets*, with *35 of those datasets becoming open and publicly available*. *All data have been anonymised* before getting ingested to the Corpus. Moreover, an *Augmentation Engine* has been developed, which can produce *augmented versions of the ingested datasets*, facilitating the Machine Learning (ML) operations for training and model evaluation. It supports around *19 augments types for video* and *43 augments types for audio* sources. The *datasets* are disseminated under a *Creative-Commons license*, while the Corpus *implementation* is going to become *open-source under an Apache 2.0 license*.

Several evaluation studies were conducted for performance, user-friendliness, security, and privacy. *Benchmarking activities* took place in two main phases due at M14 and M30 (under WP5). The Corpus *performance was good and complied with the defined requirements* set during the design phase of the project. *Three user acceptance studies* were held. The first two were performed by internal/consortium users and the third and last one by external users. The final version of the Corpus was found *ease-to-use and quite user-friendly*. The core *security evaluations* were performed with the *Assurance Platform*. The final Corpus version is *secure with no known critical vulnerabilities*. *Privacy assessments* were performed with the *SENTINEL Platform*, provided by the collaborative EU-funded project SENTINEL. Elements of self-assessment Data Protection Impact Assessment (DPIA) were performed. Pilots have also conducted their internal DPIAs before sharing any data with the project. *Privacy requirements have been properly met* for *privacy-by-design, data security, accountability, and privacy governance*. Moreover, an *Ethics Advisory Board (EAB)* was established and *verified these results*.

A roadmap has been agreed among the involved partners to *retain the Corpus operation for up to one year after the project's completion*. Also, a *joint exploitation plan* is considered. The main offerings are the *public datasets*, the *open-source implementation*, the *current MARVEL Corpus deployment*, as well as *potential new Corpus instances* sold as *Infrastructure-as-a-Service (IaaS)* or *Platform-as-a-Service (PaaS)*. Moreover, *a community of 27 academic and*

industrial groups has been formed *across 14 European countries*, with experts and practitioners *expressing their willingness to examine the use of the Corpus offerings*.

D2.5 is the final result of the tasks ***T2.3 – Incremental scheme: continuous augmentation of the dataset*** and ***T2.4 – Sharing multimodal Corpus-as-a-Service: fostering the European data economy vision in smart cities***, and the final outcome of ***WP2 – MARVEL multimodal data Corpus-as-a-Service for smart cities***, along with the related deliverable ***D2.6 – Ethics, privacy and data protection compliance management – final version*** which details the legal and privacy aspects involved in MARVEL project and the Corpus. The two deliverables contribute to the completion of the milestone ***MS8 – Long-term sustainability and commercialisation*** and the Key Performance Indicators (KPIs) ***KPI-O5-E1-1-KPI-O5-E4-1***.

DRAFT

1 Introduction

The *scarcity of publicly available datasets from smart cities poses a significant challenge for researchers and practitioners in the field of Machine Learning (ML)*. Smart cities generate vast amounts of data through interconnected devices and sensors, providing a rich source for developing and testing ML algorithms. However, due to *privacy concerns, security issues, and proprietary considerations*, many cities are reluctant to release their datasets to the public. The sensitive nature of the information collected, such as individual mobility patterns or environmental sensor data, requires careful handling to prevent misuse. Striking a balance between the potential for innovation through ML research and the protection of citizens' privacy and security remains a complex task. As a result, the *limited availability of comprehensive and diverse smart city datasets hampers the progress of ML research in urban environments*, impeding the development of robust and effective algorithms tailored to the intricacies of modern city living.

The *MARVEL project developed a platform for the gathering of data from smart city infrastructures* (i.e., surveillance systems) and the *creation of open and publicly available datasets, taking into account privacy, security, and data protection considerations*. The *objectives* were to: i) *stimulate research* on multimodal audio-visual analytics and representations and overcome the lack of publicly available processed data, ii) *collect and analyse* the nature and format of experimental data assets, and iii) *foster the European Data Economy vision* and promote free flow of data and data exchanges to support the emergence of the data markets in a smart city environment. Therefore, the *MARVEL Data Corpus* was created and released as a service [1], enabling smart cities to build and deploy innovative applications that are based on multimodal perception and intelligence.

This document is the main report for the MARVEL Data Corpus and the related activities. Technically, the Data Corpus is a *Big Data repository* that stores the *datasets that were created* during the lifespan of the project by the three pilots in the *cities of Trento and Malta* and the *Novi Sad University*, as well as the collaborative *city of Gozo* (in Malta) [2]. The repository is *deployed in a cloud infrastructure of PSNC* in Poland and is extended across several Virtual Machines (VMs). Each VM is equipped with a significant hard disk space (around 100-250 TBs). The main repository technologies include: i) the distributed file system *Hadoop* [3] – which is installed in all repository VMs and where the dataset files (video, audio, and audio-video) are actually stored; and ii) the non-SQL (NoSQL) database *HBase* [4] – which stores the metadata of each dataset and imported file, along with indices in the files in the Hadoop [5]. *Application Programming Interfaces (APIs)* have been created to ingest, update, delete, search and view/download data in this Corpus. The user *can interact programmatically or through two Graphical User Interfaces (GUIs)* – one private for the internal MARVEL users with full access privileges on all stored data, and a public one for the external users with search and view/download only functionality on the public datasets.

At first, the processed data are privately stored and are available only to the consortium members. *All data have to be anonymised*, i.e., via the MARVEL components *VideoAnony* and *AudioAnony*, before further processing and dissemination. When all data protection procedures are completed, the data owner can make a final check and choose if the data can become public or remain confidential. Moreover, the user can select from a wide list of video and audio augmenters to produce augmented versions of these datasets and facilitate the ML activities of the external users. *Augmentations utilise the libraries Keras* [6], *TensorFlow* [7], *imgaug* [8], *audiomentations* [9], and *torch-audiomentations* [10]. Thereupon, several *integration flows* have been implemented. Normally, data are *ingested after the processing by*

the anonymisation components (i.e., AudioAnony and VideoAnony) *at the Fog or Cloud layers*. Nevertheless, the *direct ingestion from stream sources* (e.g., public cameras at the edge) *has been also tested*. Also, Artificial Intelligence (AI) components can process the data and disclose video or stream snippets with specific events. These snippets are maintained by the StreamHandler component while the events are reported to other MARVEL components via a Kafka message broker. The Corpus also supports *ingestion from the StreamHandler's end*, as well as the *consumption of the reported events in the designated Data Fusion Bus (DFB) Kafka topics*. Therefore, datasets with specified events and descriptions per ingested file can be produced, not just continuous recordings.

Performance-related requirements were set during the first phase of the project under WP1. *Benchmarking* activities were conducted under WP5. All Corpus requirements *were met by the final version*. Also, two *internal evaluations* were performed by the MARVEL partners concerning *the ease of use and user-friendliness of the GUIs*. The first-round comments were addressed along with some bug-fixing and the revisited version was in general quite accepted. A *final evaluation* was performed *by external users*, who also found the Corpus easy to use and user-friendly.

Thereupon, there were performed two main assessment studies for security and privacy, respectively. *Security evaluations* were performed with the *Assurance Platform* [11]-[13] provided by STS. This platform performed a *vulnerability assessment* on the Corpus deployment on a periodic basis, disclosing known vulnerabilities of the system that have been discovered by international cybersecurity institutions (i.e., Common Vulnerabilities and Exposures (CVE) records [14]). Moreover, a *dynamic security assessment* was also conducted, utilising penetration testing elements (i.e., the use of OpenVAS). In the first iteration, it was disclosed that a vulnerable version of Hadoop had been used. A newer Hadoop version was then installed, fixing the problem. The last assessment revealed no critical security issues. *Privacy evaluations* were performed with the *SENTINEL Platform* [15]. SENTINEL is a collaborating EU-funded project, which has developed a platform to assess the *General Data Protection Regulation (GDPR) compliance* [16] of Small and Medium-sized Enterprises (SMEs). The first iteration revealed that there were not sufficient procedures to record the activity of the system and respond to GDPR-related requests by data owners. Therefore, the monitoring and logging elements were enhanced, as well as the procedures to handle requests from data owners (i.e., pilots) or other entities. These elements were proven very significant, when at the last phase of the project there were requests from Italian privacy authorities that were successfully answered, concerning the access and the amount of data from the Trento use cases.

All *public datasets* are disseminated under a *Creative Commons (CC) license*. Moreover, the *database application* that was developed, will be also available as an *open-source project under an Apache 2.0 license* for anyone who wants to build a similar repository for his/her own. Apart from licensing, the dedicated task *T2.5* examined ethics, privacy, and data protection compliance issues. An *Ethical Advisory Board (EAB)* was established to audit the MARVEL approach and provide feedback. Details can be found under the deliverable D2.6 (due at M36).

A *roadmap* has been decided for the *support of the Corpus for up to one year after the end of the project*. PSNC has agreed to continue providing the infrastructure during this period. Henceforth, *Letters of Intent (LOIs)* will be signed by several partners denoting their will to support this activity. FORTH and STS (which implemented the core Corpus services) will provide technical support. The pilots MT, GRN, and UNS will retain their public datasets. Also, PSNC provides the infrastructure under a *Service Level Agreement (SLA)* concerning the

availability of the resources. Utilisation of the infrastructure is *continuously monitored* via *Zabbix agents*, while the *Assurance Platform* is also utilised for the monitoring of security and privacy principles.

To further support the sustainability of the Corpus, a *joint exploitation plan* has been established between the involved partners. The Big Data Repositories can be sold as *Infrastructure-as-a-Service (IaaS)* or *Platform-as-a-Service (PaaS)*. There is a decent number of *open datasets*, while the database application can be offered as an *open-source project*. Apart from the main products, services for *technical support, consultancy, and training* can be offered. The *security and privacy monitoring* with the Assurance Platform can be also included as an additional service in the main IaaS or PaaS. More details can be found under the deliverable D7.6 (due at M36).

Regarding the *data volume*, many *obstacles and restricting factors* were raised during the course of the project. The most important one was the fact that it was chosen to *utilise video streams of lower resolution, high compression, and low frames per second (fps) to meet the real-time processing requirements of the rest of MARVEL operations* (i.e., anonymisation, AI processing and inference of incidents, data management and transmission, etc.). Therefore, datasets of lower size in TBs were produced, compared with the initial vision when the proposal was written, where it was calculated that streams of very high resolution and fps would have been processed. At the end of the project (M36), the Corpus reached a *size of around 1.1 PBs*. Nevertheless, the *ingestion and augmentation of datasets can be continued for up to one year* as the computational resources will still be available.

Although this is lower than the initial KPI for 3.3PBs, the collected datasets and implemented technologies were proven to be very useful for end-users and help in building a *community of academic and industrial organisations*, which provided very positive feedback and are willing to use it for their activities. A short list of these organisations includes:

- Academic and Research partners
 - Universities – Utilisation within ML courses and student thesis
 - Hellenic Mediterranean University (HMU) (<https://hmu.gr/en/home/>)
 - Aristotle University of Thessaloniki (AUTH) (<https://www.auth.gr/en/>)
 - University of Piraeus (UNIPI) (<https://unipi.gr/unipi/en/>)
 - European University Cyprus (EUC) (<https://euc.ac.cy/en/>)
 - University of South Wales (USW) (<https://www.southwales.ac.uk/>)
 - Metropolitan College (AMC) (<https://www.mitropolitiko.edu.gr/en/>)
 - Mediterranean College (MEC) (<https://medcollege.edu.gr/>)
 - Aarhus University (AU) (<https://international.au.dk/>)
 - University of Novi Sad (UNS) (<https://www.uns.ac.rs/index.php/en/>)
 - Tampereen University (TAU) (<https://www.tuni.fi/en/>)
 - Research Institutes – Use by ML researchers
 - Foundation for Research and Technology – Hellas (FORTH) (<https://forth.gr/en/home/>)
 - Italian National Research Council (CNR) (<https://www.cnr.it/en/>)
 - Fondazione Bruno Kessler (FBK) (<https://www.fbk.eu/en/>)

- SMEs and Industrial partners
 - SMEs
 - Techpro Academy (<https://www.techproacademy.gr/en/>) supported by Deloitte, Netcompany Intrasoft, ALTAIR, and DataScouting – Utilised within courses for professional training programs
 - DRAXIS (<https://draxis.gr/>)
 - UBITECH (<https://ubitech.eu/>)
 - Raven Cybersecurity (<https://ravensec.eu/>)
 - AEGIS IT Research (<https://aegisresearch.eu/>)
 - *GreenRoads (GRN)* (<https://www.greenroads.ai/>)
 - *ZELUS* (<https://www.zelus.gr/>)
 - *Sphinx Technology Solutions (STS)* (<https://www.sphinx.ch/>)
 - Industry
 - SWORD (<https://www.sword-group.com/>)
 - *Intrasoft International (INTRA)* (<https://www.netcompany-intrasoft.com/>) – Store data streams
 - *Infineon (IFAG)* (<https://www.infineon.com/>)
 - *ATOS* (<https://atos.net/en/>)
 - *audEERING GmbH (AUD)* (<https://www.audeering.com/>)
 - *Poznan Supercomputing and Networking Center (PSNC)* (<https://www.psnc.pl/>)

1.1 Purpose and scope

This deliverable details the implementation aspects of the MARVEL Data Corpus and summarises the related activities that were performed throughout the lifespan of the project. The Corpus is planned to remain in use for up to one year from the end of the project. D2.5 is also acting as a reference for the potential external users of the current Corpus deployment, the ingested public datasets, as well as the use of its open-source implementation. A guide for the use of the public GUI is provided, along with a detailed description of the related APIs and data models that someone can use to interact with the MARVEL Corpus programmatically. Also, the sustainability and exploitation aspects are provided for everyone who wants to support and promote our vision.

1.2 Relation to other work packages, deliverables, and activities

This work is conducted under the *WP2 – MARVEL multimodal data Corpus-as-a-Service for smart cities*. The deliverable *D2.5 – Corpus-as-a-Service specifications and business continuity* is the final outcome of the tasks *T2.3 – Incremental scheme: continuous augmentation of the dataset* and *T2.4 – Sharing multimodal Corpus-as-a-Service: fostering the European data economy vision in smart cities* and presents the materialisation of the MARVEL Data Corpus. The deliverable *D2.6 – Ethics, privacy and data protection compliance management – final version* of the collaborative task *T2.5 – Ethics, privacy and data protection compliance*, documents the legal and GDPR compliance aspects of the

MARVEL framework and the stored data. The two deliverables constitute the final results of WP2 and contribute to the completion of the milestone *MS8 – Long-term sustainability and commercialisation* and the Key Performance Indicators (KPIs) *KPI-O5-E1-1-KPI-O5-E4-1*.

Also, this work took input from task *T2.1 – Collection and analysis of MARVEL experimental distributed data assets*, concerning the specifications and data models that were defined during the first design phase of the project, while technical aspects and data management were supported by the components that were developed under the task *T2.2 – Data management and distribution*. For the reader's reference, a first dataset's specification is recorded in D2.1 [17], while the initial and final versions of the data management tools can be found in D2.2 [18] and D2.4 [19], respectively.

The main design goal of the Corpus was to store the datasets that were produced by the pilots under *WP6 – Real-life societal experiments in smart cities environment* and the related tasks and use cases. Requirements and main architecture were specified under *WP1 – Setting the scene: Project set up*. The Corpus infrastructure was provided and tested within the activities of *WP5 – Infrastructure Management and Integration* and the underlying tasks.

Finally, the outcomes of this work support the exploitation and communication activities that were held under *WP7 – Exploitation, sustainability, and business continuity*, as the MARVEL Data Corpus is one of the Key Exploitable Results (KERs) of the project which is examined under the Innovation Radar [20].

1.3 Structure of the deliverable

The deliverable is structured as follows. *Section 2* provides a high-level overview of the MARVEL Data Corpus, including design goals, technical implementation and deployment details, analysis and evaluation trials, legal aspects, as well as the lessons learned throughout the project lifespan. *Section 3* details the technical information for the Corpus implementation and the deployment of the Big Data repository. *Section 4* presents the available user interfaces (graphical and programmatic). *Section 5* describes the offered functionality of producing augmented versions of the public datasets. *Section 6* summarises the datasets that are available in the Corpus (both the public datasets and the ones that are maintained privately for the MARVEL partners). *Section 7* reviews the various analyses that were conducted concerning the evaluation of performance, user acceptance, as well as security and privacy properties. *Section 8* sketches the main exploitation strategy and the actions towards building a community of experts from industry and academia who will make use of Corpus. *Section 9* concludes the document and summarises the overall results.

Finally, there are annexes that describe in detail: *i*) the Data Model that was used for datasets (*Annex 1*), *ii*) the template for the Letters of Intent (LOI) that were signed by the partners that they will support the service operation after the end of the project (*Annex 2*), *iii*) the three user satisfaction studies that were performed (2 by internal users and 1 by external ones) (*Annex 3*), and *iv*) a detailed justification concerning the lower total size of the data, the difficulties that we face, and the mitigation actions that were performed (*Annex 4*).

2 Overview of the MARVEL Data Corpus

This Section provides a summary of the overall contributions related to the MARVEL Data Corpus. These include: *i*) the main ideas behind its implementation; *ii*) the status of the Corpus until M36 in terms of infrastructure and collected datasets; *iii*) the main technologies that were utilised for its development; *iv*) the evaluation studies that were conducted for its performance, user-friendliness, security, and privacy; *v*) the legal aspects concerning usage, licensing of data, and its open-source implementation; *vi*) the exploitation strategy for this KER; as well as *vii*) the lessons learnt throughout the project.

2.1 Goals, Objectives, and Vision

Smart cities, gaining attention from various stakeholders like governments, policymakers, municipalities, industry, and researchers, have evolved into intricate systems driven by extensive data collection facilitated by a wide variety of IoT sensors and devices scattered across urban landscapes. These devices, including cameras, microphones, and temperature sensors, amass vast datasets reflecting the daily activities of citizens. However, the substantial challenges emanating from the collection of diverse, large-scale datasets necessitate strategies for extracting meaningful insights that can translate into commercial benefits. To tackle this, there is a demand for enhanced, precise predictions and analytics. ***Conventional methods prove inadequate due to the real-time, voluminous, and high-throughput nature of the incoming data.*** Therefore, the solution lies in adopting novel methodologies, techniques, and tools that can effectively extract valuable knowledge and commercial value from this deluge of information, addressing the problem of extreme-scale data analytics.

Smart cities have become “data engines”. However, there is a ***lack of available data sources to research and industrial communities***, which could make use of such data and create new solutions and business models that could benefit the citizens (e.g., assist police or other authorities, enhance urban planning and ecological policies, etc.). Henceforth, the ***goals of MARVEL*** can be outlined as follows:

- Develop a privacy-conscious solution to unveil valuable insights aimed at enhancing the quality of life.
- Facilitate decision-making through event detection and situational awareness in smart city environments.
- Overcome technological silos to foster integrated and collaborative solutions.
- Validate the system in real-world scenarios for practical applicability.
- Contribute significantly to extremely large-scale audio-video datasets, thereby supporting the European data-driven economy.

MARVEL’s ***vision*** is to gather, analyse, and perform data mining on multi-modal audio-visual data streams originating from a Smart City, and thereupon, provide decision-makers with insights to enhance the quality of life for citizens and improve the services offered to them. All while adhering to ethical standards and privacy constraints in a responsible AI fashion.

Data are gathered from various sources (i.e., surveillance cameras, microphones, and drones), are anonymised, annotated, and augmented throughout the Edge-to-Fog-to-Cloud (E2F2C) architecture of MARVEL [2], and are stored in the form of datasets in the MARVEL Data Corpus repository. Then, the datasets are either disseminated publicly as open datasets or are utilised internally by MARVEL’s AI/ML components to enhance their training procedures. Figure 1 depicts the main concepts of the Data Corpus solution.

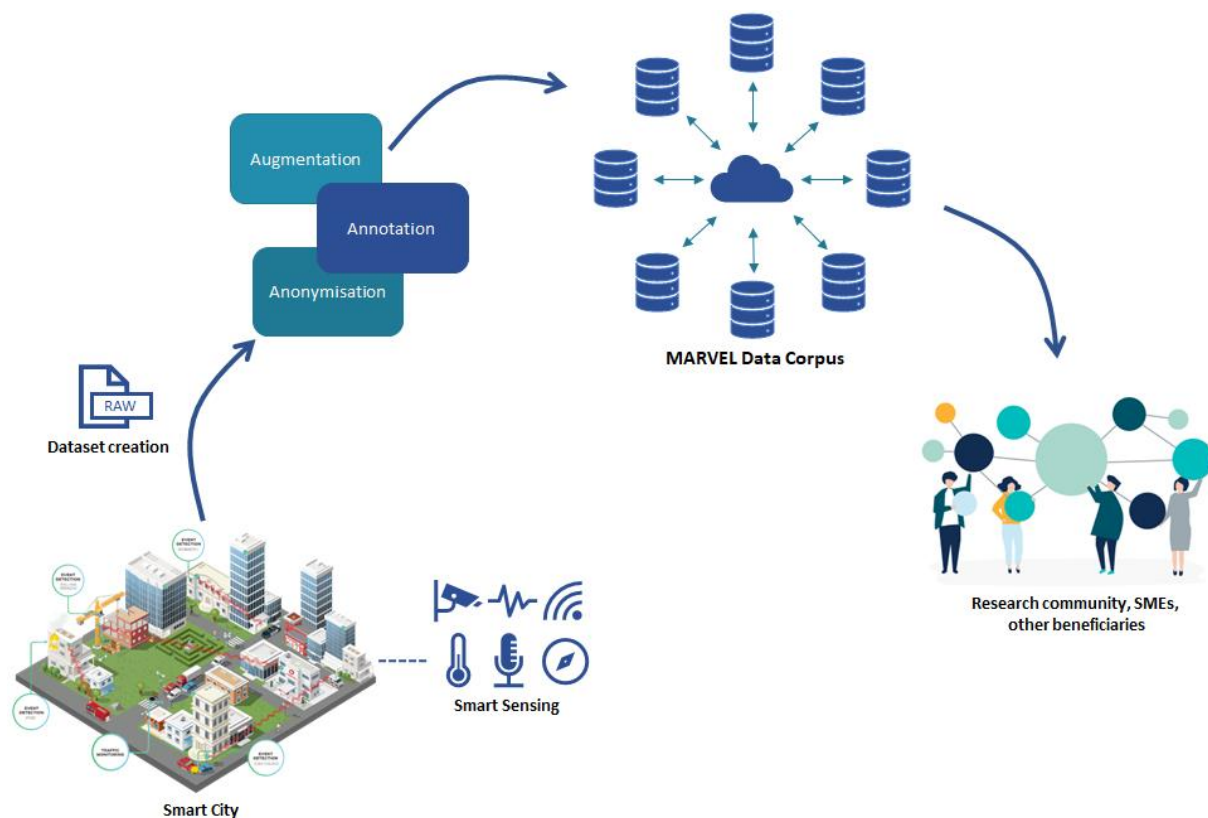


Figure 1. MARVEL Data Corpus main concepts

The *core objective* of Corpus is to “*foster the European Data Economy vision and create new scientific and business opportunities by offering the MARVEL Data Corpus as a free service and contributing to BDVA standards*” (GA – Objective 5). The Corpus success is measured with the completion of 4 KPIs, as outlined in Table 1.

Table 1. KPIs evaluation

KPI	Description	Evaluation
KPI-O5-E1-1	More than 3.3PBs of data made available through a Corpus-as-a-Service	The current size of the Corpus is around 1.1PBs. Although the size is lower than the targeted level, the produced datasets were quite valuable and were appreciated by the community of external users that was established. The main restricting factor for a higher size was the fact that we had to process video of high compression and low fps to react processing requirements for real-time operation of the AI/ML components. See Annex 4 (Section 14) for the full justification regarding the accomplished volume.
KPI-O5-E2-1	Release SLAs and consider all the relevant aspects, namely accessibility, operability, managing streaming and network, legal considerations, security,	PSNC provides the Corpus infrastructure under an SLA. Moreover, the Corpus utilisation is continuously monitored via Zabbix agents, while the security and privacy aspects are monitored via the Assurance Platform. FORTH and STS administrate the developed application and respond to system breakdowns and user requests. PN is also providing

	privacy, and technical concerns	legal consultancy and periodic monitoring of the privacy and ethical landscape. See subsection 2.5 for more details.
KPI-O5-E3-1	More than 5 SMEs used the Corpus	Through several info days and dissemination/communication events, a community of around 25 organisations has been established that seems willing to utilise the work done. This includes academic partners from universities, colleges, and research institutes, as well as industries and 8 SMEs. See Section 8 for more details.
KPI-O5-E4-1	Maintain the corpus for at least one year after the end of the project	A sustainability plan has been agreed upon among the involved partners to retain the Corpus operation for at least one year after the end of the project. Thereupon, PSNC will continue providing the Corpus infrastructure, FORTH and STS will provide maintenance to the developed application and potential technical support to external users, while the pilots MT, GRN, and UNS will retain the public datasets accessible. See Section 8 for more details.

2.2 Current Status and Available Data

Several data flows were integrated with the Corpus, ingesting data from the three piloting environments, as well as an additional city of Gozo. *Datasets from all use cases* were produced (4 for MT, 4 for GRN, 2 for UNS) [2], while the *Gozo scenarios* are related to urban planning and traffic management as for GRN.

Around 64 datasets were produced. From them, 29 datasets from MT are confidential for the moment, until the privacy authorities of Italy come to a final decision concerning the dissemination of this data. The *total size* of data that were collected until M36 is around **1.1 PBs**.

The public datasets can be accessed through the following link: <https://datacorpus.marvel-project.eu/>. Moreover, users who want to contribute to the MARVEL Data Corpus with their own datasets can request access and become part of the community (email to: marvel-info@marvel-project.eu), as in the case of Gozo.

The covered use cases and the defined datasets are presented in Section 6.

2.3 Main Technologies Used

The Data Corpus is built in the infrastructure provided by PSNC. The main technologies that were used for the development of our solution, are: i) the distributed file system of **Hadoop** for the implementation of the core repository of data [3]; ii) the NoSQL database **HBase** for the implementation of the database application with indexing to the files in Hadoop [4]; and iii) the **Angular 2.0** framework [21] for the development of user-friendly GUIs.

Under task T2.3, the main augmentation techniques were deployed as well. Five main augmentation libraries were considered: **Keras** [6], **TensorFlow** [7], **imagaug** [8], **audiomentations** [9], and **torch-audiomentations** [10].

Moreover, several integration flows with other MARVEL components were materialised. All data are anonymised before getting ingested to the Corpus. The main components for this are *VideoAnony* and *AudioAnony* for the anonymisation of video and audio, respectively. Apart from the direct integration with the anonymisation tools in the E2F2C, there was also an integration with the *SmartViz/DMT* through the *StreamHandler* and the message broker *Kafka* (DFB component).

The main functionality of these technologies is summarised in the following paragraphs.

2.3.1 Hadoop repository

Hadoop [3] is a widely used framework in Big Data applications due to its scalability, fault tolerance, and cost-effective storage through the Hadoop Distributed File System (HDFS). Its parallel processing capabilities accelerate data tasks, and its flexibility allows handling both structured and unstructured data. The comprehensive ecosystem includes tools for processing, storage, querying, and real-time processing. As an open-source framework with a large community, it encourages innovation and customisation. The optimisation of data locality, proven reliability, and successful implementations by major enterprises further solidify Hadoop's position in managing and processing massive datasets within the dynamic landscape of Big Data technologies.

2.3.2 HBase database and HDFS

HBase [4] is another powerful choice for Big Data applications, particularly in scenarios demanding real-time, random access to large datasets. Its column-oriented NoSQL database design provides rapid data retrieval, making it suitable for applications requiring low-latency access, such as time-series data or real-time analytics. HBase seamlessly integrates with the Hadoop ecosystem [5], leveraging the scalability and fault tolerance of HDFS, and offers automatic sharding for horizontal scalability. Its schema flexibility accommodates dynamic data models, crucial in scenarios where data structures evolve over time. HBase's ability to handle vast amounts of sparse data efficiently, coupled with strong consistency guarantees, makes it an optimal solution for applications demanding high-speed, scalable, and reliable data storage and retrieval within the realm of Big Data.

The Corpus database implementation is detailed in subsection 3.2.

2.3.3 Angular 2.0 for GUI development

Angular 2.0 [21] is a robust and versatile framework that brings several advantages to the development of modern GUIs. Its use of TypeScript, a statically typed superset of JavaScript, enhances code maintainability and catches errors during development, ensuring a more stable application. Angular's modular architecture facilitates code organisation and scalability, making it well-suited for large and complex projects. The two-way data binding and dependency injection mechanisms streamline the development process, enabling developers to create dynamic, responsive, and data-driven interfaces more efficiently. Additionally, Angular's extensive ecosystem, including a rich set of libraries and a vibrant community, provides ample resources and support. Its commitment to a component-based structure promotes reusability and maintainability, contributing to a more structured and efficient GUI development process. Overall, Angular stands out as a powerful framework for crafting modern and feature-rich graphical user interfaces.

The developed Corpus GUIs with Angular 2 are presented in Section 4.

2.3.4 Augmentations

Keras [6] offers significant advantages in the augmentation of video and audio data, particularly due to its simplicity, flexibility, and integration with deep learning frameworks like **TensorFlow** [7]. Keras provides a high-level, user-friendly interface, enabling developers to easily implement complex neural network architectures for video and audio data processing. Its modular design and extensive documentation make it accessible to both beginners and experienced practitioners. Keras supports a wide array of data augmentation techniques, crucial for enhancing the variety and quality of datasets, which is particularly beneficial for training robust models in video and audio processing tasks. With its seamless interoperability with TensorFlow, Keras leverages the strengths of deep learning libraries, ensuring efficient computation and scalability. Overall, Keras simplifies the augmentation process, fostering innovation and experimentation in video and audio data applications within the deep learning domain.

Also, **imgaug** [8], a powerful image augmentation library, can be effectively employed for augmenting video data by extending its capabilities to handle individual frames within video sequences. While not originally designed for video, **imgaug**'s diverse set of transformations, including geometric and colour augmentations, can be applied frame-wise to enhance the diversity of video datasets. By incorporating **imgaug** into video pre-processing pipelines, users can introduce variability and robustness to their video datasets, aiding machine learning models in better generalisation and performance across different scenarios.

The **audiomentations** [9] is a Python library for audio data augmentation. It is useful for machine learning and runs on CPU. The library supports mono audio and multichannel audio and can be integrated into training pipelines in, for instance, Keras, TensorFlow, or PyTorch. It has also helped people get world-class results in Kaggle competitions [22]. Similarly, **torch-audiomentations** [10] has been inspired by **audiomentations** and provides audio data augmentation in PyTorch. This library supports CPU and GPU (CUDA), batches of multichannel (or mono) audio, and cross-platform compatibility. It is designed with high speed and high-test coverage in mind.

The developed augmentation strategy for Corpus is presented in Section 5.

2.3.5 Anonymisation

All data that are ingested in the Data Corpus have been anonymised in advance [23]. This processing is materialised through the MARVEL components **VideoAnony** and **AudioAnony** for video and audio anonymisation, respectively. The main operation of the two components is documented in the deliverable **D3.2 – Efficient deployment of AI-optimised ML/DL models – initial version** [24] (due at M18), while the final versions are updated in **D3.6 – Efficient deployment of AI-optimised ML/DL models – final version** [25] (due at M30).

VideoAnony processes raw video streams from CCTV cameras, employing facial and car plate anonymisation through blurring techniques. Its objective is to obscure identified faces and license plates in the live video feeds obtained from Closed-Circuit Television (CCTV) cameras at each pilot site. Anonymisation is executed using a range of methods, starting from traditional image processing approaches like blurring and advancing to sophisticated techniques such as Generative Adversarial Networks (GAN) for face-swapping, which have been developed within the MARVEL project. **VideoAnony** acquires the incoming video stream through Real-Time Streaming Protocol (RTSP). It utilises the YOLOv5 detector [26] for face and car plate detection, fine-tuned with public datasets and pilot-provided annotations. Detected regions of interest are then blurred. The initial project phase tackled challenges like pose preservation and

varying face sizes in CCTV videos, while the subsequent phase focused on reducing the model's computational complexity through component replacement and weight quantisation.

AudioAnony performs real-time processing of audio streams to identify sensitive audio content, primarily present in speech segments. This module collaborates with MARVEL's Voice Activity Detection (VAD) module, ensuring waveform modifications occur only when speech is detected. The chosen audio segments undergo anonymisation using signal-processing-based techniques. It relies on fundamental signal processing, generating a new waveform by modifying associated poles computed from Linear Predictive Coding (LPC) coefficients on a frame basis. This process shifts the formant positions, altering the spectral envelope and voice characteristics while preserving speech content. Formant shifting is controlled by a single parameter, the McAdams coefficient. In the absence of speech, the unmodified waveform is transmitted. Processed segments, whether anonymised or not, are sent to later stages of the MARVEL audio-visual pipeline for consumption by various audio AI components. The module also records analysis details by forwarding event boundaries to the related Message Queuing Telemetry Transport (MQTT) broker, subsequently stored in the Elastic Search database and utilised by SmartViz.

The integration of Corpus with the anonymisation components is presented in subsection 3.3.

2.3.6 AI inference and Data management

After anonymisation, data streams can be processed by MARVEL's AI components, before getting ingested to the Corpus. The main AI inference pipeline is visualised by the *SmartViz* and *Decision-Making Toolkit (DMT)*. The MARVEL framework processes the data streams in real-time and by applying training ML models specific events or anomalies can be recognised. Once such an incident is recognised, the related stream part is stored locally by the *StreamHandler* component in a form of a file (e.g., video or audio file). Also, information concerning this incident is getting published in DFB Kafka topics, and other components can consume this information. Therefore, these files from the StreamHandler part can be ingested to the Corpus along with their descriptions denoted in the *Kafka* topics.

SmartViz and DMT have been implemented under task *T4.4 – MARVEL's decision-making toolkit*. Their final versions have been presented in the deliverable *D4.6 – MARVEL's decision-making toolkit – final version* [27]. SmartViz functions as the user interface (UI) for the DMT, offering a means to visualise detected events, anomalies, and alerts generated from the analysis of data from the AI components that are processed at the various layers depending on the use case [28], [25]. Comprising various visualisation tools, SmartViz supports exploratory analysis by presenting interactive data representations. This allows end users to engage with the data, acquire a thorough understanding, delve into detailed information, and recognise patterns and correlations.

StreamHandler and Kafka are part of the overall data management components [29]-[30] that have been defined under the activities of task *T2.2 – Data management and distribution*. Their final versions have been described in the deliverable *D2.4 – Management and distribution Toolkit – final version* [19].

The Kafka broker [31], a core component in the Apache Kafka distributed streaming platform, acts as a centralised hub for managing the communication and data flow between producers and consumers within a Kafka cluster. It facilitates the seamless, fault-tolerant, and scalable exchange of streaming data by persistently storing and managing the topic-based message logs. Serving as a highly durable and distributed commit log, the Kafka broker ensures reliability and fault tolerance in real-time data streaming applications. Its distributed architecture allows for

horizontal scaling and high throughput, making it a robust and efficient solution for building resilient and scalable data pipelines.

The StreamHandler is positioned in the fog and receives audio/video (AV) data streams from active sources like CCTV cameras, network-enabled microphones, AudioAnony, and VideoAnony instances via RTSP. During initialisation, it queries the AVRegistry's REpresentational State Transfer (REST) API to discover active AV sources and their details. In operation, StreamHandler consumes AV RTSP streams and segments them into binary documents based on predefined time intervals for persistent storage. It archives these files and provides a REST API for SmartViz to request the transmission of AV data from specific sources and time points. Upon such requests, StreamHandler compiles a unified stream corresponding to the requested timeframe, generates a link to the binary file, and transmits it to SmartViz. Additionally, StreamHandler's REST API accepts requests from SmartViz to transmit on-demand AV data via RTSP, requiring parameters such as the original AV source's ID and the absolute start time.

The integration of Corpus with these technologies is presented in subsection 3.3.

2.4 Implementation and Evaluation Stages

2.4.1 Implementation roadmap

Based on the GA, the main task **T2.4 – Sharing multimodal Corpus-as-a-Service: fostering the European data economy vision in smart cities** that was related to the **Data Corpus implementation** was meant to start at M20. Nevertheless, main elements had been deployed under **T2.3 – Incremental scheme: continuous augmentation of the dataset** activities with the development of the Minimum Valuable Product (MVP) due at M12 (i.e., Hadoop and HBase in a single VM along with main Data Corpus APIs to ingest data) and with the 1st MARVEL prototype due at M18 (i.e., clustered version of Hadoop/HBase in several VMs acting as data nodes). The full operation of the Corpus (i.e., all APIs, GUIs, security monitoring, and infrastructure) was integrated in the 2nd MARVEL prototype due at M30. The main data flows, which had been deployed to ingest data from the piloting environments, continue operation until M36. Meanwhile, minor updates and bug fixing also took place.

2.4.2 Performance and benchmarking

For the **performance evaluation**, two main benchmarking trails took place at M14 and M30, which were documented in the deliverables **D5.2 – Technical evaluation and progress against benchmarks – initial version** [32] and **D5.5 – Technical evaluation and progress against benchmarks** [33], respectively. In each iteration, a series of ingestions from several parallel data streams were performed, with performance metrics being measured. In general, the throughput of the Corpus services was within the targeted limits that had been set during the design phase of the project, and no data loss or other malfunction or failure was recorded. The benchmarking results are detailed in subsection 7.1.

2.4.3 User acceptance

Moreover, three **user evaluation** studies took place to assess the ease of use of Corpus and user acceptance. The **first and second assessments** were **by internal MARVEL users**. Fruitful feedback was collected during the initial trial, while the second one revealed that a decent user-friendliness level has been achieved. Thereupon, a **third assessment** was performed **by external users**. These users are experts from organisations which seem willing to utilise the Corpus offerings and help us establish a MARVEL Data Corpus community to sustain operation after the end of the project. This study also verified that a good level of user acceptance had been

finally accomplished. The user evaluations are presented in Section 7.2 and Annex 3 of this document.

2.4.4 Security assessment

Security evaluations were performed with the STS's *Assurance Platform* [11]-[13]. A series of assessments were conducted during the various implementation phases of the project.

The platform consistently conducted periodic *vulnerability assessments* on the Corpus deployment, systematically revealing any known vulnerabilities within the system, as documented in Common Vulnerabilities and Exposures (CVE) records [14]. Furthermore, *dynamic security assessments and penetration testing* were performed, such as analysis with OpenVAS. During the initial evaluation, it was identified that an outdated and vulnerable version of Hadoop had been implemented. Promptly, a more recent version of Hadoop was installed to address and rectify the issue. In subsequent assessments, the system displayed an absence of critical security issues, signifying the effectiveness of the implemented measures in enhancing the platform's security posture.

Thereupon, *Assurance Profiles* were deployed for the continuous assessment of the deployed security and data protection mechanisms on the Corpus. These profiles use event captors to gather information about the running system and reason about its *CIA aspects*. The results are visualised in a web interface where the MARVEL user can monitor the Big Data infrastructure in terms of security elements.

The Assurance Platform functionality, as well as the performed security assessments, are detailed in subsection 7.3.

2.4.5 Privacy assessment

Comprehensive privacy evaluations were conducted using the *SENTINEL Platform* [15], a collaborative project funded by the EU dedicated to assessing GDPR compliance specifically tailored for SMEs. Among others, this platform includes elements of automated *Data Protection Impact Assessment (DPIA)*. The initial assessment brought to light deficiencies in the existing procedures for recording system activities and responding to GDPR-related requests from data owners. Consequently, significant enhancements were made to the monitoring and logging components, along with the refinement of procedures to effectively address requests from data owners, such as the project pilots, and other relevant entities. This iterative process ensures ongoing alignment with GDPR standards and reinforces the platform's commitment to privacy and data protection.

The main GDPR-compliance aspects [16] of the Data Corpus were examined under task **T2.6 – Ethics, privacy and data protection compliance** and the detailed results are documented in the deliverable **D2.6 – Ethics, privacy, and data protection compliance management – final version**. Privacy requirements have been properly fulfilled, e.g., for privacy-by-design, data security, accountability, and privacy governance. Moreover, Pilots performed internal DPIAs before processing any data. An *EAB* was also formed and verified the results.

The SENTINEL Platform functionality and the performed privacy assessments are detailed in subsection 7.4.

2.5 Legal Aspects

Several legal concepts are involved in the dissemination and sustainability of the Data Corpus, including the elements of:

- Licensing of data;

- Letters of Intent (LOI) from the partners that they will continue supporting the Corpus operation for up to 1 year after the completion of the project;
- Service Level Agreement (SLA) for the Corpus' infrastructure; and
- Open-source licensing of Corpus repository implementation.

To make the produced datasets open and permit public sharing, the pilots provide their data based on related a *Creative Commons (CC) license*. External users of the Corpus can obtain the data and use them for research or other purposes. The MARVEL Corpus has to be acknowledged in that case.

Based on the GA, there was the obligation to continue the Corpus operation for at least 1 year after the project's completion. Apart from PSNC which provides the infrastructure, no other partner was specifically mentioned for the achievement of this goal. However, as PSNC just provided the computational resources but did not participate in the Corpus implementation itself, several partners have to support this operation as well. Therefore, partners that were involved in the Corpus materialisation, had to sign a *Letter of Intent (LOI)* to state their willingness to do so. This includes: i) FORTH and STS which made the core development and will provide technical support and administration of the Corpus during this period; and ii) the pilots MT, GRN, and UNS which act as the data owners and will have to retain their datasets open to the public.

Moreover, to enhance the users' acceptance and promote the adaptation of our solution for potential commercial use and exploitation, the *SLA-driven operation* of the Corpus was established. Thus, PSNC provides the infrastructure under an SLA. Zabbix agents have been deployed to monitor the Corpus resources and verify that the implemented services and applications are accessible. Furthermore, the Assurance Platform is supporting this feature by deploying Assurance Profiles and verifying that security and privacy principles are met.

Another offering of the Data Corpus activities is to disseminate the main repository implementation upon the framework of Hadoop/HBase under an open-source license. The *Apache 2.0 license* was found to be the most appropriate for our goals. Thereupon, the implementation will be shared as an open project in GitHub. It is considered that this feature will be of significant value for the overall exploitation of the MARVEL project due to the popularity of Hadoop/HBase in Big Data implementations, promoting also our activities towards open science.

2.6 Exploitation strategy

The main offering of the Data Corpus is a Database for multimodal Big Data. "Corpus" is an umbrella term for several concepts, like:

- the infrastructure (PSNC);
- the implementation of the Database application and services upon this infrastructure (STS, FORTH);
- the security and privacy monitoring modules (STS);
- and the data ingested in the current installation (MT, GRN, UNS).

Joint exploitation plans have been defined among the involved partners. The potential use models for the diverse components within the Data Corpus implementation encompass various strategies for commercialisation and collaboration.

- For *IaaS*, exemplified by PSNC, a commercialisation approach involves selling the Big Data Repository based on storage size, with offerings ranging from 0.5PB to larger capacities (e.g., offerings for 0.5PB, 1PB, 2PB, etc.).
- In the case of the *Database application*, led by STS and FORTH, an exploitation strategy is adopted by providing services freely under an open-source license, fostering open-source initiatives. Moreover, commercialisation avenues include offering technical support, training, and consultancy services on-demand or through subscription models.
- *PaaS*, involving PSNC, STS, and FORTH, combines the Repository infrastructure and the Database application, mirroring IaaS commercialisation models with added services, such as the STS Assurance Platform.
- *Data* from the MARVEL Corpus, contributed by MT, GRN, UNS, and Pilot Owners, is made available under the CC-BY-NC license, facilitating free use for academic and research purposes. Commercialisation possibilities involve on-demand agreements with pilot owners for data-driven business models, aligning with specific use cases like urban planning or traffic monitoring.

Overall, the Data Corpus is anticipated to be publicly accessible before the project's conclusion, with PSNC maintaining the infrastructure for up to a year post-project. STS and FORTH may offer technical support, and pilots commit to retaining data accessibility during this period.

The Corpus exploitation strategy has been presented under the *WP7 – Exploitation, sustainability, and business continuity* activities.

2.7 Lessons Learned

During the course of the MARVEL project, there were significant obstacles that were raised, and we had to overcome them to implement and support the operation of the Data Corpus. Unforeseen effort, technical difficulties, and legal considerations are some of them.

One main issue that was faced, was the fact that it *was underestimated the effort that would be required to implement, safeguard, and manage a Big Data infrastructure with sensitive data* that has to *become publicly available*. As the whole MARVEL Platform was supposed to reach up to a Technology Readiness Level (TRL) 5 (aka “technology validated in relevant environment” from the HORIZON 2020 definition [34]) based on the GA, it was planned to produce a relatively “simple” component where data would just be stored and later be retrieved by some external users.

However, as the processed data were privacy-sensitive, a data owner and its privacy authority would be very reluctant to share citizens’ data in a “research repository” with low technological maturity. Therefore, *concrete security controls* had to be built. These included a wide range of controls for security, privacy, and data protection, as well as continuous monitoring mechanisms, extensive logging and logfiles’ management, and regular assessments of security and privacy elements and properties.

Also, as the need for *GDPR compliance* was foreseen as a necessity, *relevant procedures* had to be set, *along with the technical solutions that would support them*. For example, to respond to a request from a municipality’s privacy authority concerning the processing of its citizens’ data, we had to set a communication channel for GDPR requests, establish procedures and responsibilities for the proper and timely handling of these requests, as well as the technical implementations to collect information – such as the volume of data for the specific municipality, the duration of this data, who had access from the consortium and the potential

external users, the type of access and processing that these users did, timeline-related information, etc.; and form a report that would be included in the formal response of the request. This imposed additional technical, managerial, and administrative efforts to support the proper operation of the Corpus.

Moreover, as the volume of data was continually getting increased as well as the number of sources that ingested data to the Corpus, several components were *facing performance issues progressively*. When an issue across the integrated data flow was fixed, another element would become the new bottleneck after some period. Therefore, there was a periodic need to improve performance for the bottlenecks that were revealed, until the current efficient state was finally reached. These included efforts to make computations more efficient, parallelise processing and manage parallel components (e.g., set procedures to semi/automatically continue operation after restarts or breakdowns), and add computational and communicational resources.

Despite all this progress and the delivery of a quite mature and safe solution, *ethical concerns still remain* about recording citizens' activity and producing AI/ML algorithms to process it. In spite of the AV data being anonymised (e.g., faces and voices blurred), and the inference data not including any personal data, almost near the end of the project, officials from the Municipality of Trento became very concerned about the overall approach and began re-evaluating their position. Privacy authorities rethinking the legal basis of this research and start assessing the various elements more consciously. The aforementioned implemented procedures assisted in answering their focused requests and helped in clarifying issues and resolving potential misunderstandings. The municipality will make its final decisions after the end of the project (M36) and will determine whether it will keep the MT datasets public or not.

Henceforth, apart from the additional technical functionality for security, privacy, and performance, it was required to devote significant effort in administrating all these elements. Moreover, the activation of these procedures for the case of the Municipality of Trento put even more stress on the personnel that was working on the Corpus tasks. Nevertheless, the proactive development of all those elements ensured the proper handling of the situation.

3 Technical Description of the MARVEL Data Corpus Database

This Section details the technical aspects of the Data Corpus. These include: *i*) the acquired infrastructure; *ii*) the deployment of the Corpus elements; *iii*) the implementation of the database application; *iv*) the integration with the rest of the MARVEL architecture; and *v*) the main security and privacy controls.

The primary objective of MARVEL's Data Corpus is to enhance open science and open data initiatives by fostering research and innovation, thereby advancing the benchmarking of edge, fog, cloud, and AI technologies. Provided as a service, the Data Corpus facilitates the development and implementation of innovative applications by smart cities, leveraging multimodal perception and intelligence as outlined in Pillar I of the project in GA.

3.1 Infrastructure & Deployment

The MARVEL Data Corpus has been deployed upon the cloud infrastructure of PSNC that could be extended and reach a total storage space of 3.3PBs. Initially, a low set of storage resources was provided. As data were getting ingested to the Corpus, these resources were extended every time the storage usage level was reaching 80%. At its final stage within M36, the datasets in the Corpus reached a volume of around 1.1PBs. Nevertheless, PSNC can provide several PBs upon request (even beyond the targeted 3.3PBs).

The main counterparts of the Corpus repository are: *i*) the distributed file system Hadoop [3], and *ii*) the No-SQL database HBase [4]. This combined Hadoop/HBase framework is a popular solution for Big Data repositories [5].

The Corpus repository is deployed across two large region servers of PSNC (DCW and BST), where 8 VMs were constructed as it is detailed below. The implementation of the core Corpus repository was made on the Hadoop Distributed File System (HDFS). Figure 2 depicts the resource monitoring panel of the HDFS at an early deployment stage with 5 VMs and low-resource utilisation.

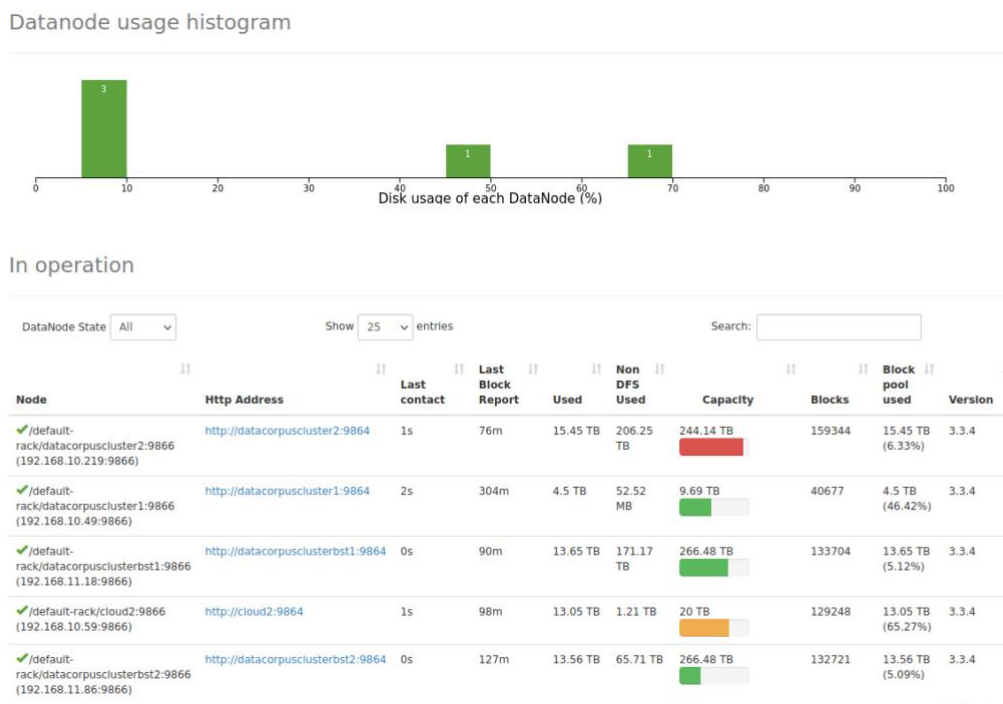


Figure 2. The Data Corpus VMs status

One VM acts as the Master Node, where all the main components of the Hadoop/HBase framework are installed. The rest VMs act as Data Nodes of the HDFS, which are operated by this master node. Figure 3 illustrates the main building blocks of the repository implementation. The components defined for Data Corpus architecture include:

- One Region Server.
- The Region Server talks with a Master Node on HDFS.
- The Master Node splits the data into portions (slaves – Data Nodes), depending on the size of the data.
- Zookeeper manages the HBase.
- Queries (accessing) to the database are made through Ambari

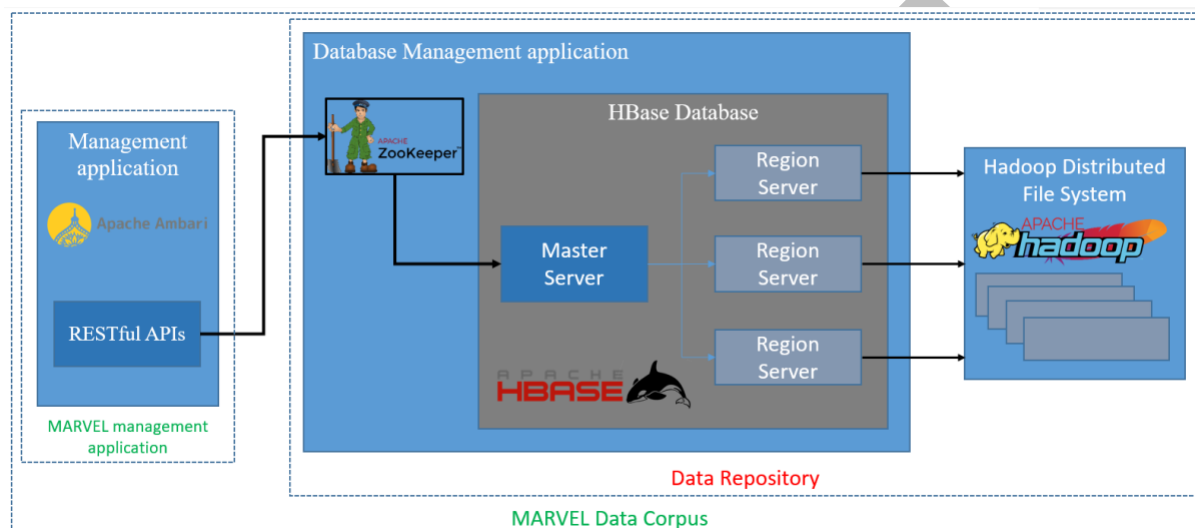


Figure 3. Main building blocks of the core Data Corpus repository implementation

A specific number of VMs were allocated at each time-point to deploy the Hadoop/HBase cluster. The storage backends for these VMs were improved to meet capacity and high throughput parameters by using the clusters with significant storage volumes. Based on PSNC's experience, it was suggested to create VMs with hard disk spaces that will not exceed 250TBs. This was the best trade-off for the creation of a PB-scale repository of good performance. Larger hard disks could be deployed in the given infrastructure (e.g., 500TBs or even 1TB), but this would result in lower performance (e.g., larger seek, read, and write times).

Secure Shell (SSH) keys are deployed in every machine and their communication is protected, offering node authentication as well as confidentiality and integrity on transit. The networking of all VMs is materialised via internal 1Gbs Ethernet links.

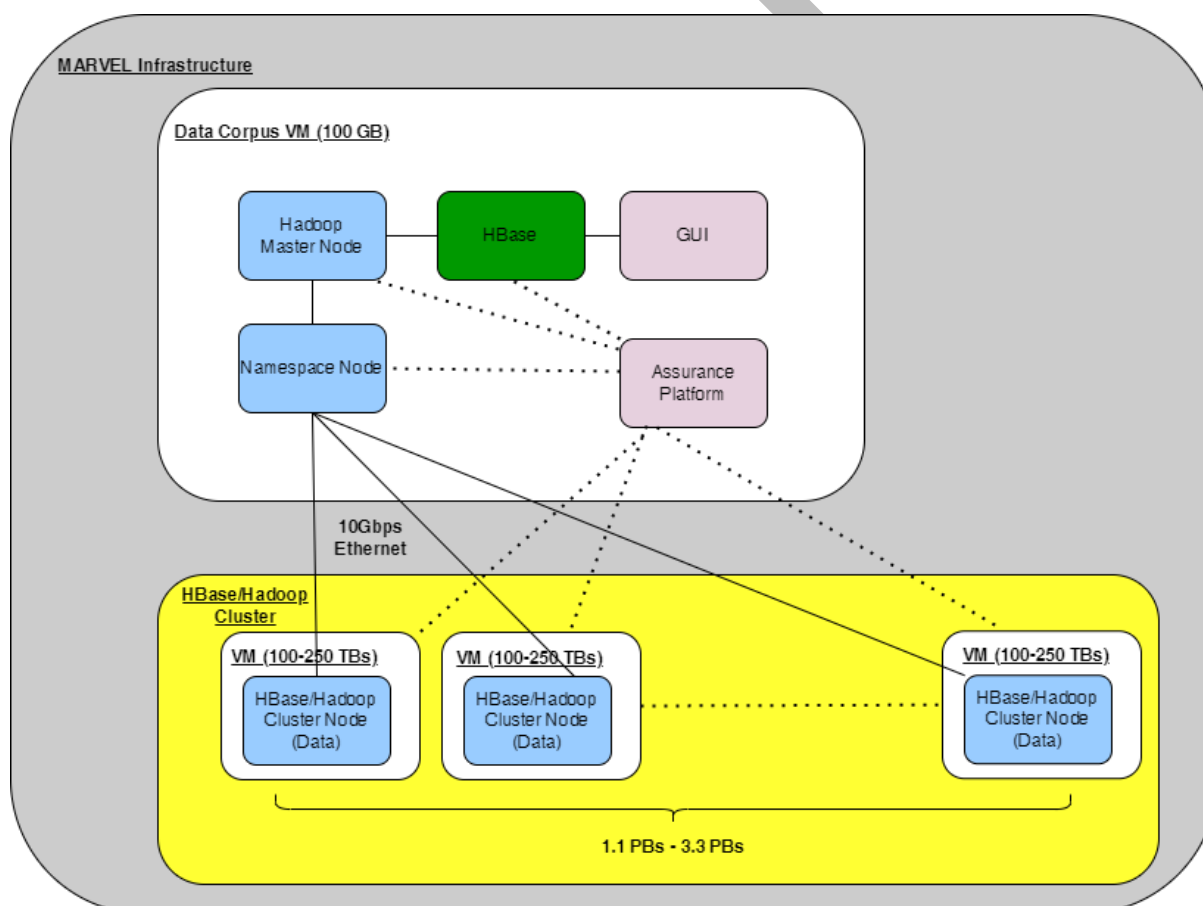
Also, the Assurance Platform was deployed to monitor the availability of HBase and Hadoop in the Master Node. The Assurance Platform deploys custom event captors or Beats of the Elasticsearch, Logstash, and Kibana (ELK) stack to collect and process information concerning the running system.

The current installation comprises of VMs presented in Table 2. The first entry is related to the Mater Node, which was given mode CPUs and Memory in order not to face performance issues during congested periods.

Table 2. Configuration of the Data Corpus VMs

No.	# of VCPUs	Memory	Volume	Operating system
1	16	32GB	100TBs	Ubuntu Server 20.04 LTS
2	8	16GB	100TBs	Ubuntu Server 20.04 LTS
3	8	16GB	100TBs	Ubuntu Server 20.04 LTS
4	8	16GB	100TBs	Ubuntu Server 20.04 LTS
5	8	16GB	100TBs	Ubuntu Server 20.04 LTS
6	8	16GB	150TBs	Ubuntu Server 20.04 LTS
7	8	16GB	250TBs	Ubuntu Server 20.04 LTS
8	8	16GB	250TBs	Ubuntu Server 20.04 LTS

As mentioned before, the provided infrastructure is scalable and can be extended, simply by adding new VMs as Data Nodes under this HDFS. A configuration is needed for each VM (i.e., install HDFS elements and SSH keys), as well as to add the new node in the Master Node's configuration file. Figure 4 illustrates the deployment of the core elements of the MARVEL Data Corpus and the potential extension up to 3.3.PBs.

**Figure 4.** Infrastructure and Deployment view on several VMs

3.2 Implementation of Database Application and Services

Upon the main Hadoop/HBase installation, the MARVEL Corpus database application was built. The database schema is based on the Data Model that was defined for the specification of the MARVEL datasets (see Annex 1).

The Corpus consists of a list of datasets. Each Dataset has several general descriptive meta-data, including:

- Dataset Name and ID
- Description
- Keywords
- Start date of recording
- Provider/Owner of the dataset (aka. the 3 pilots)
- The involved use case.
- Data category (i.e., audio, video, or audio-video)

There are also some *generic technical meta-data*, including:

- Recording device
 - Device ID
 - Latitude and Longitude
- Duration of recorded snippet files
- Annotations
 - Annotation software
 - Annotation ontology
- Augmentation method (if applicable)
- Anonymisation method (if applicable)

Moreover, there are several *technical descriptive meta-data based on this dataset category*.

These include:

- Audio
 - Bitrate
 - Sampling
 - Number of channels
- Video
 - Resolution
 - Frames per second
- Audio-video
 - All the above

Most importantly, each dataset contains a list of *Snippets*, which are the main data themselves. Ordinarily, the Snippets are subsequently recorded files from a single device (e.g., camera,

microphone, or drone) of a similar duration. The Snippets details are presented in the following subsection.

The **Snippets** are the main data of the Corpus. Each Snippet represents a specific recorded file and is part of a single Dataset. A Snippet is a conceptual structure that can contain up to three of the following data files:

- The **main recorded file**, usually stored in an ordinary audio (e.g., MP3, WAV, etc.) or video (e.g., MP4, AVI, MKV, WMV, etc.) format. In some cases, there can be a compressed file in a ZIP format with several *augmented versions* of the original file (e.g., for different levels of brightness, with added noises, or filters simulating various weather conditions).
- The **annotation file**, usually stored in a simple TeXT file extension (TXT) or structured Extensible Markup Language (XML) format and contains annotations for the related main recorded file.
- Similar to the annotations, the **inference results file** in an XML format, which has been produced by the MARVEL's AI components, after processing the main recorded file of the Snippet.

As with the Datasets, each Snippet has a set of descriptive metadata, including:

- Snippet Name and ID
- Publication date
- Captured events in this snippet
- Start and end timestamps
- Duration

Therefore, **two tables are defined in the HBase**, one **for the dataset entries** and one **for the snippet entries**, respectively. For each dataset, an entry is created in the dataset table, storing the descriptive metadata. Similarly, for each snippet, an entry is created in the snippet table. In each snippet entry, there is a field denoting the key to the related dataset entry.

The **database backend was implemented in JAVA**. Thereupon, **APIs** were developed for the main database functionality of:

- *Insert* datasets and snippets
- *Update* existing dataset and snippet entries
- *Delete* existing entries
- *Search* for entries based on their metadata
- *View* ingested entries
- *Download* a snippet file
- *Metrics*
 - Total datasets and snippets
 - Datasets/snippets per contributor (i.e., MT, GRN, and UNS)
 - Datasets/snippets per data type (i.e., audio, video, and audio-video)
- *Internal APIs* (not directly utilised by users), e.g.:
 - List the snippets of a dataset
 - Count public or private datasets
 - Count original or augmented datasets

In total, **around 35 APIs** were implemented. The APIs are operating under the **user authentication mechanisms** developed for the overall MARVEL Platform.

Moreover, graphical interfaces were implemented for the front-end to facilitate the Corpus use. **Two GUIs were developed in Angular 2.0** [21]. The *private GUI* is designed for the MARVEL users and gives access to the full Corpus functionality. The *public GUI* is similar but allows only the functionality to search, view, and download. The GUIs are presented in detail in Section 4.

A MARVEL user can ingest data via the private GUI. He/she fulfils the descriptive fields for the dataset entry and its snippets. The user can browse his/her file system and select the snippet files that will be ingested. The user can either select one-by-one the files or a folder containing all dataset's snippets.

However, for the project's objectives, there was also the need to ingest high volumes of data programmatically. Therefore, an **ingestion tool** was also **developed in JAVA**, providing similar functionality. The user provides the dataset's description in a JSON format (similar to the one presented under Annex 1) and specifies the folder with the snippets. This ingestion agent will read the contents of the folder one-by-one, will create the related snippet entries in the database and upload the underlying files. The ingestion with this tool can be stopped and resumed by the user. Also, the ingestion can be automated and activated periodically or when new files/data are available in the folder. This is the main ingestion mechanism that was utilised during the course of the project. The various integrations with the other MARVEL components and data streams are detailed in the following subsection.

3.3 Integration with the MARVEL Platform

Figure 5 illustrates the final MARVEL architecture. This Section outlines the integration flows that were implemented. Data can be ingested: *i)* right after the execution of the anonymisation components, *ii)* when the AI components identify some event, or *iii)* by the user via a web interface. Except from anonymised data, **augmented versions** can be also ingested.

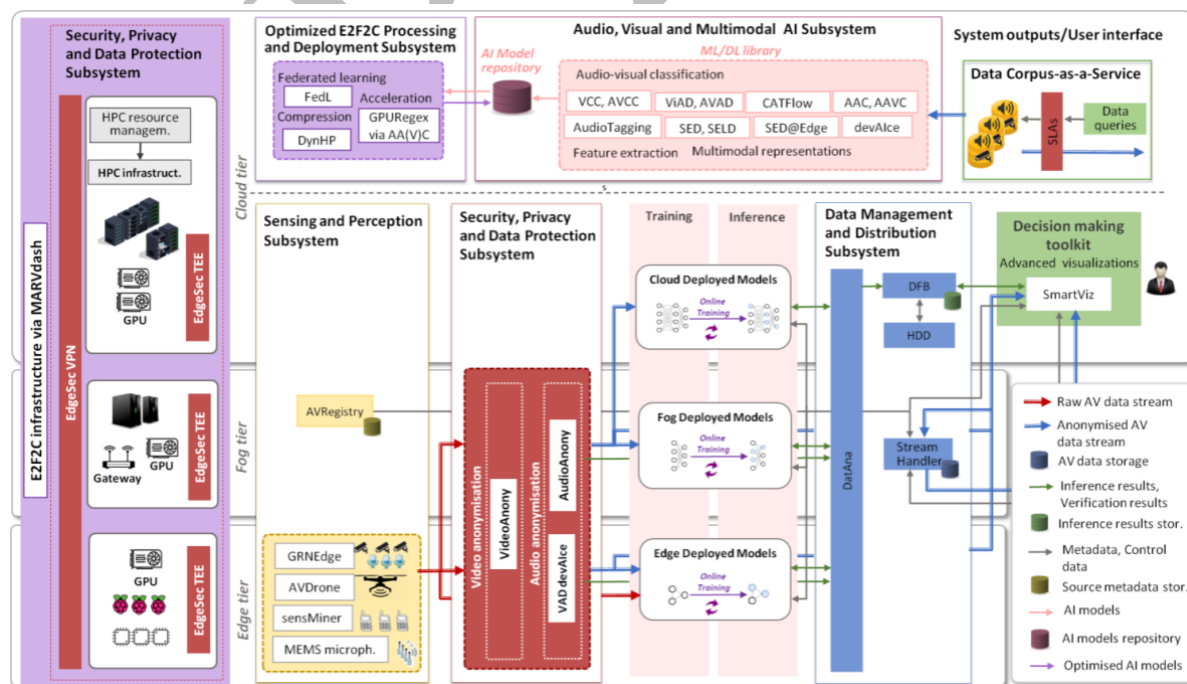


Figure 5. MARVEL conceptual architecture

3.3.1 Ingestion of anonymised piloting data from Edge-Fog-Cloud

One main data flow that was implemented, involves the ingestion of data to the Corpus right after their anonymisation. The anonymisation components across the E2F2C, like *VideoAnony* and *AudioAnony*, take as input a video/audio stream from a device (e.g., surveillance camera or microphone), anonymise the data, and produce files of constant duration (e.g., every 1 hour) with the new anonymised version. The files are stored in a relevant folder. Therefore, the user can create a related dataset entry in the Corpus and initiate a process to ingest these files, as the snippet files of the dataset. The user fills in the dataset's metadata and then runs an ingestion tool, which reads one-by-one the files from the folder and calls the Corpus APIs to ingest the data.

This method proved very useful as the integration was simple, could be automated, and did not require from several components to be in a stable and running mode to ingest data. In general, as many MARVEL components were built from scratch, such components were usually unavailable due to development, integration, or maintenance tasks. While this integration flow required only a stream source to be up and running, the anonymisation, and the final ingestion to the Corpus. These elements were relatively stable throughout the project, and once such a flow from a device was set, it could run for several weeks and months without interaction.

The drawback of this method is that the produced datasets were lacking rich metadata. Except for the initial description that was provided by the user, no additional events or annotations were recorded for the ingested snippets.

3.3.2 Ingestion of real-time data with identified events from the AI inference pipeline via StreamHandler

Another integration flow involves the ingestion of data from the *StreamHandler* as well as inference results produced by the *AI inference pipeline* and communicated to the Corpus via a *DFB Kafka* message broker. This integration is depicted in Figure 6.

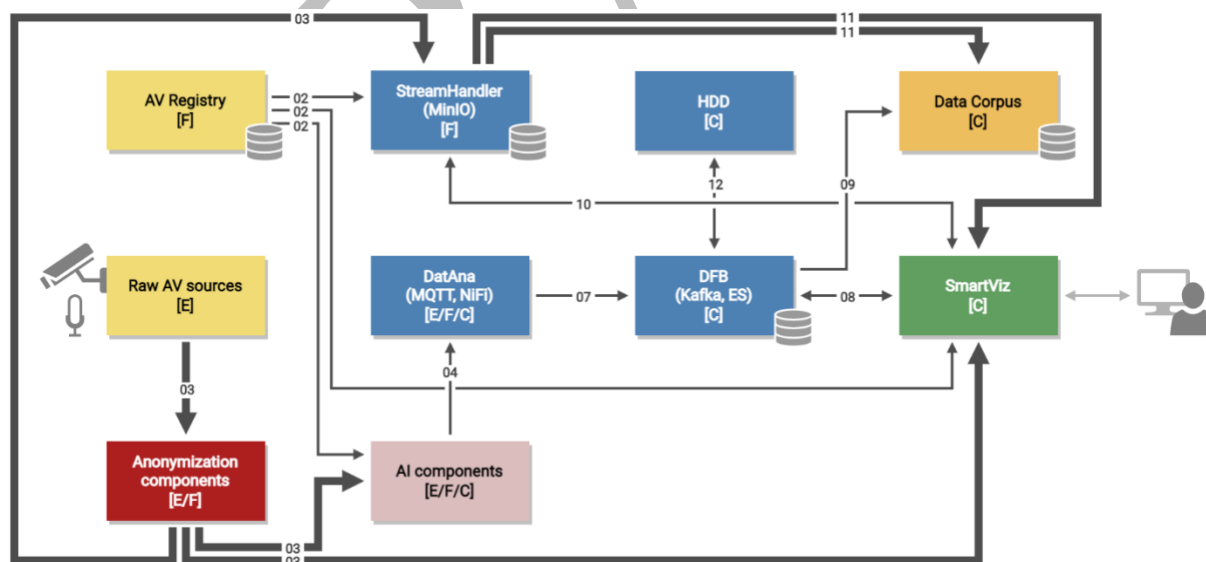


Figure 6. Integration of Data Corpus with the StreamHandler and AI inference pipeline

Here, when events were detected by the AI Inference pipeline in a processed stream, via Kafka it notified the StreamHandler about the involved stream frames (with information about *cameraId*, *startTime*, *endTime* and *timestamp* if *startTime*, *endTime* are not present), and the StreamHandler produced snippet files with the related frames. The files were stored in different

folders for each different stream source (i.e., streaming device). An example of these folders' organisation for the GRN cameras is depicted in Figure 7.

```

gre1@grn1og2:/mnt/store/files/private/manf/minio_data/minio-segments$ ls -la
total 29464
drwxr-xr-x 13 root root      303 Jun 27 05:47 -
drwxr-xr-x  5 gre1 gre1       98 May 24 09:03 ..
drwxr-xr-x  2 root root       10 Jun 27 05:47 bin_#
drwxr-xr-x  2 root root       10 Jun 21 10:02 Cam-GRN-CCTU-01
drwxr-xr-x  2 root root       10 Jun 21 10:02 Cam-GRN-CCTU-02
drwxr-xr-x  2 root root       10 Jun 21 10:02 Cam-GRN-CCTU-03
drwxr-xr-x  2 root root       10 Jun 21 10:02
drwxr-xr-x 46 root root 17911808 Oct 27 00:03 Cam-GRN-UA-01
drwxr-xr-x 25 root root 1134592 Oct 27 07:44 Cam-GRN-UA-01-Audio
drwxr-xr-x 55 root root 2605056 Oct 27 07:34 Cam-GRN-UA-02
drwxr-xr-x 49 root root  864256 Oct 27 07:44 Cam-GRN-UA-02-Audio
drwxr-xr-x 55 root root 2818048 Oct 27 07:36 Cam-GRN-UA-03
drwxr-xr-x 49 root root  888832 Oct 27 07:44 Cam-GRN-UA-03-Audio
drwxr-xr-x  2 root root   610304 Oct 24 06:57 generated-media
gre1@grn1og2:/mnt/store/files/private/manf/minio_data/minio-segments$

```

Figure 7. StreamHandler's output on folders

Also, AI modules publish the event in dedicated Kafka topics for other components to consume it. Thereupon, the abovementioned ingestion tool could be activated in each of these folders, creating new datasets that contain more interesting snippet files where some identified activities were taking place. Moreover, the event descriptions could be consumed by the Corpus from a dedicated Kafka topic and be included in the inference results file of each relevant snippet. This *inference result's data model* is presented in the following code sample in a JavaScript Object Notation (JSON) template.

```

{
  "frame": "<The frame's timestamp>",
  "predictedAnomaly": "<True or False>",
  "imagePaths": "<The path where the frame's image can be found, e.g., path/frames/0.jpg>",
  "audioPaths": "<The path where the frame's audio can be found e.g., path/frames/0.wav>",
  "densityPaths": "<The path where the frame's density can be found>",
  "id": "<Unique identifier for this inference result, e.g., AVAD_Event_2022-04-11T06:40:10.063Z>",
  "type": "<The result's type, e.g., Anomaly>",
  "cameraID": "<The camera's unique MARVEL identifier, e.g., GRN-2>",
  "dateProcessed": "<The date where this data has been processed>",
  "name": "<The name of the stream, e.g., AVAD-2022-04-11T06:40:10.063Z>",
  "MLModelId": "<The file name of the ML model that produced this result, e.g., AVAD-v1>",
  "category": "<The result's category, e.g., Anomaly1>"
}

```

This ingestion produced more fruitful results, but the datasets were smaller than in the first case where all information was recorded. Another drawback was the fact that all these components had to be up and running and get synchronised to produce correct results. Due to the several development, testing, and revision cycles that were conducted throughout the project, this was not the main ingestion flow that was deployed. In most cases, it was fully instantiated during some benchmarking or other demonstration events and trials with test data being processed to check that the outcomes were correct. Then, the automated ingestion was stopped, as it was consuming significant resources for each streaming source and makes difficult the continuation of the technical work and the development plan. Nevertheless, the flow was tested several times

and could be set to run automatically in an actual system where the implementation would be in a mature state without requiring further development and updates.

3.3.3 Ingestion of augmented data

Apart from the original data, an *Augmentation Engine* was developed (see Section 5) under the Data Corpus. The user can produce augmented versions of the captured data in an attempt to facilitate the ML expert and enhance ML operations for training and evaluation. In general, the user can select a wide range of audio and video augmenters and apply them to the original files. The Augmentation Engine exports the new files in a designated folder. Then, he/she can create new datasets with those files as their snippets. The ingestion is performed via the tool as in the previous cases.

Several usage strategies were tested. At first, it was tested to create one new dataset for each augments with specific parameter(s) (e.g., apply the brightness augments with a brightness level equal to '50'). However, this created many datasets that were not easy to be managed by a user.

Therefore, after several iterations and consolidations with ML experts and practitioners, an augmentation strategy was defined. By default, we will apply a set of specific augmentations for a dataset (augmentation profiles). Then, the original file and all of its augmented versions will be zipped into a file, which will be the actual data that are ingested for this snippet. Therefore, this process was automated for the datasets created for Gozo. In these augmented datasets, each snippet is comprised of the relevant ZIP file. This approach was considered quite user-friendly by external users.

3.3.4 Ingestion of data from the MARVEL users' Web Interface

The final integration flow involves the user's interaction and the ingestion of datasets via the *private GUI* that was available for the MARVEL users. Here, the user utilises the Web interface to *upload data through Internet*. The user starts by inputting the descriptive information for a new dataset. Then, he/she can either upload each snippet and each file (main snippet file, annotation file, and inference result file) manually *one-by-one* or select a *folder* with all snippet files and upload them automatically (in a similar fashion as with the ingestion tool). Figure 8 depicts the successful creation of a new dataset with 5 snippet files from this private GUI.

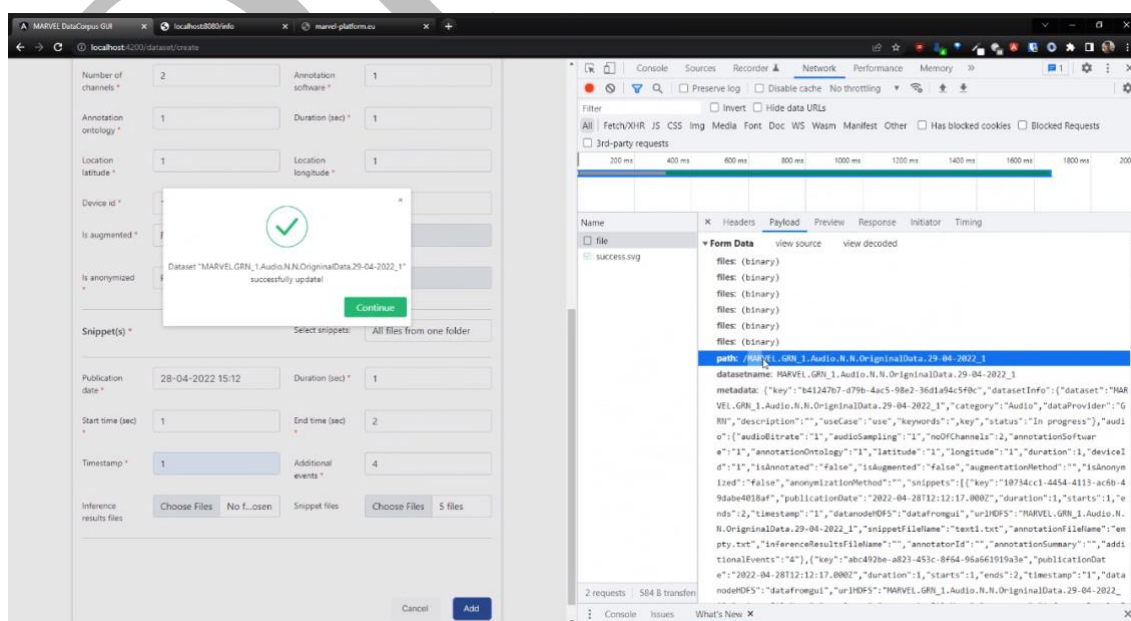


Figure 8. Create dataset from private GUI

As mentioned before, this ingestion flow was mostly used for testing, debugging, and demonstration purposes. Almost all data in the Corpus were ingested with the use of the ingestion tool via one of the abovementioned ingestion options.

3.4 Main Security and Privacy Aspects

Several security and privacy controls have been implemented to safeguard the Data Corpus infrastructure and its data. All *data are anonymised* in advance before getting ingested in the Corpus to protect the citizens' privacy. Also, all data are *protected in transit with TLS/SSL*, providing confidentiality and integrity. *User management* is implemented. The users are *authenticated* before gaining access and the policy of 'need to know' is applied, assigning only the *minimum required usage privileges* to each one. The Hadoop system supports *replication of data* to *retain availability* in case some of the Data Nodes become unavailable, while PSNC is operating *automated backup* solutions for the provided infrastructure.

Server *hardening practices* have been applied to all Corpus VMs. Among others, these include application of updates/upgrades, deletion of services or applications that are not in use, deletion of default accounts, user authentication with SSH keys, installation of anti-virus/anti-malware software, and setting firewalls.

Moreover, software suites are installed to *monitor and administrate* the running system. Zabbix agents have been deployed to monitor the *resource utilisation and system status*. The Assurance Platform is additionally deployed to monitor *security and privacy-related principles and metrics*. This Platform also acts as an *Intrusion Detection System (IDS)* that alarms the administrator when a security/privacy assurance profile is getting violated. *Logging mechanisms* have been also activated for the various Corpus components, keeping records of the performed operations and actions. The Assurance Platform can utilise ELK components (i.e., Beats) to process these log files.

Direct interaction with the infrastructure is *available only in the local network* or via *Virtual Private Network (VPN)* access. *Proxy servers* are used to hide the infrastructure from direct access, scanning, and analysis through Internet. Proxy rules are implemented to permit access to public datasets/snippets from the public GUI, without external user authentication. Authentication from the *MARVEL Platform Single Sign-On (SSO) mechanism* is required to access the private GUI and the related functionality by a MARVEL user.

Finally, several *security and privacy evaluations* were performed during the various development phases to verify that an acceptable level of protection is achieved. Security and vulnerability assessments were supported with the Assurance Platform, while privacy assessments were conducted with the use of the SENTINEL Platform. These procedures are detailed in the subsections 7.3 and 7.4, respectively.

4 User Interfaces

This Section describes the *user interfaces of the MARVEL Data Corpus*. These include: *i*) a **public web interface** where external users can access the publicly available datasets (view, search, download); *ii*) a **private web interface** for technical partners and data contributors with full access and functionality (insert, update, delete, view, search, download); and *iii*) relevant **public and private APIs** for anyone who wants to access the Corpus programmatically.

A related **user tutorial** is also available in a Portable Document Format (PDF) form on the main MARVEL website.

4.1 Public Web Interface

The Corpus stores a collection of datasets that were produced during the project's lifecycle by the three pilots (MT, GRN, UNS). The public website where external users can access the open datasets is available at the link: <https://www.marvel-project.eu/marvel-data-corpus/>.

4.1.1 Introductory Page

The first time that a user enters the Corpus webpage from a browser, it shows an introductory page of the Corpus, along with the terms of use and other legal aspects (Figure 9-Figure 11). The user has to accept these terms of use to proceed.

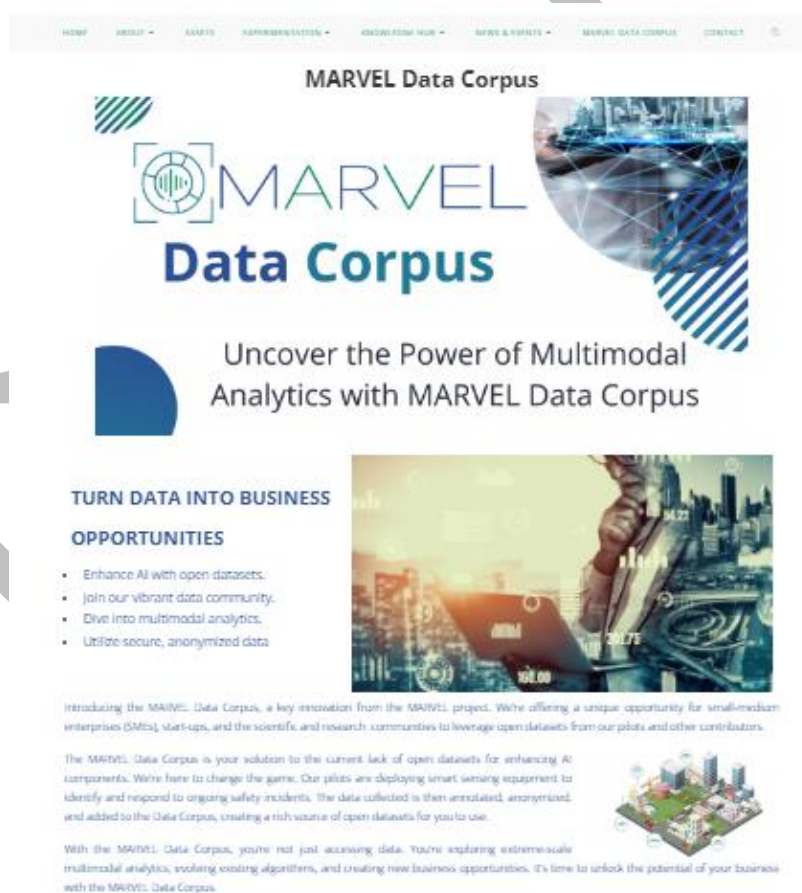


Figure 9. GUI – Introductory Page 1-1

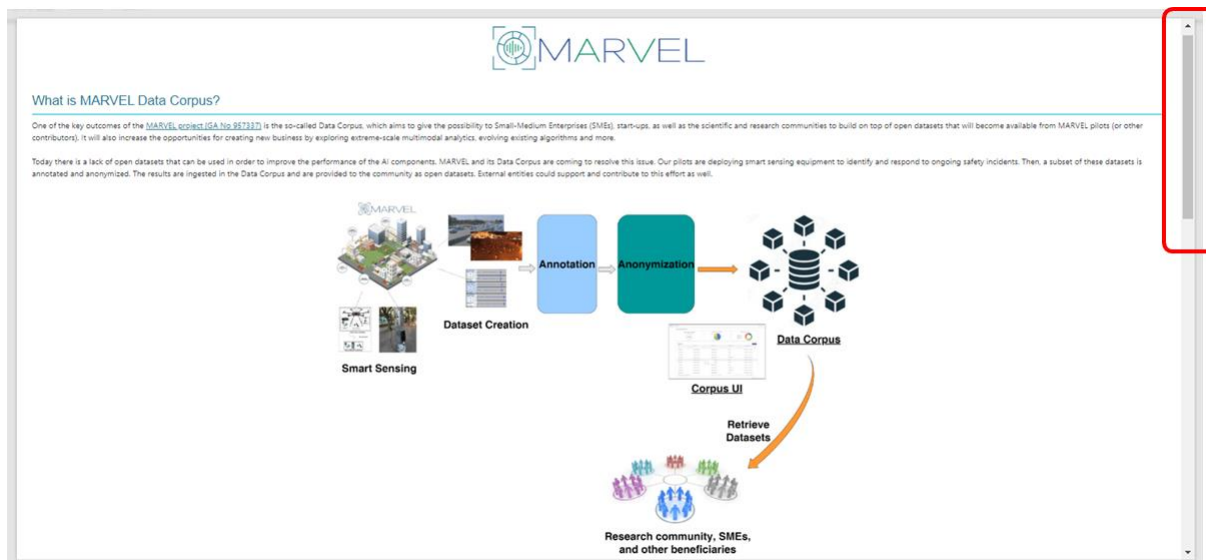


Figure 10. GUI – Introductory Page 1-2

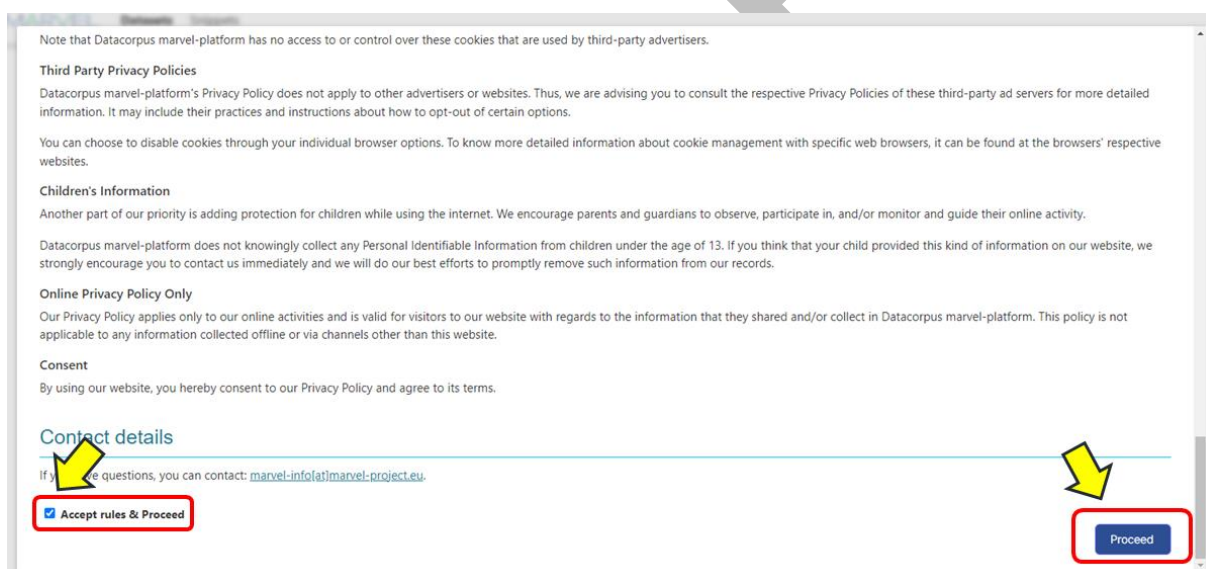


Figure 11. GUI – Introductory Page 1-3

4.2 Datasets view

Upon entering the main Corpus web page, the user sees the Datasets general view (Figure 12). This includes:

- Figures that aggregate information concerning the available Datasets.
 - Total number of Datasets
 - Datasets per Category (i.e., audio, video, or audio-video)
 - Datasets per Provider (i.e., MT, GRN, or UNS)
- The core Datasets Table

The user can: *i*) navigate through the table to view all Datasets, *ii*) filter the Datasets based on the included column fields (i.e., Dataset Name, Provider, Use Case, Category, and Ingestion date), or *iii*) search datasets based on their meta-data (Figure 13).

Once clicking on a Dataset entry in the table, the detailed Dataset view page is revealed (see next subsection 4.3).

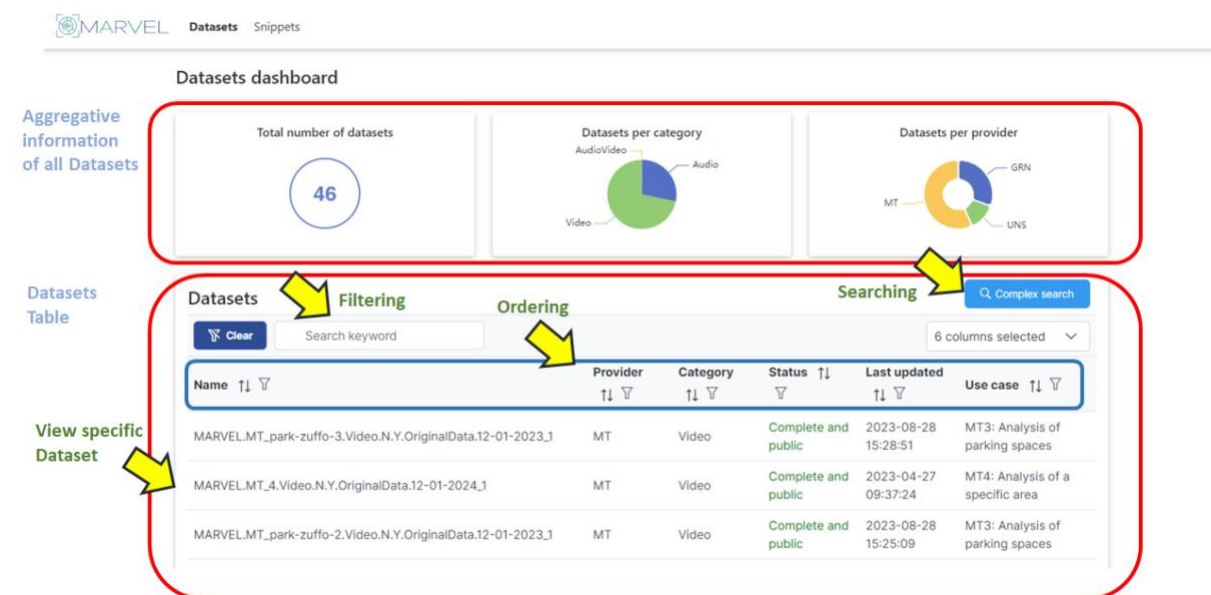


Figure 12. GUI – Datasets View 1-1 – Datasets’ Aggregated data and Table

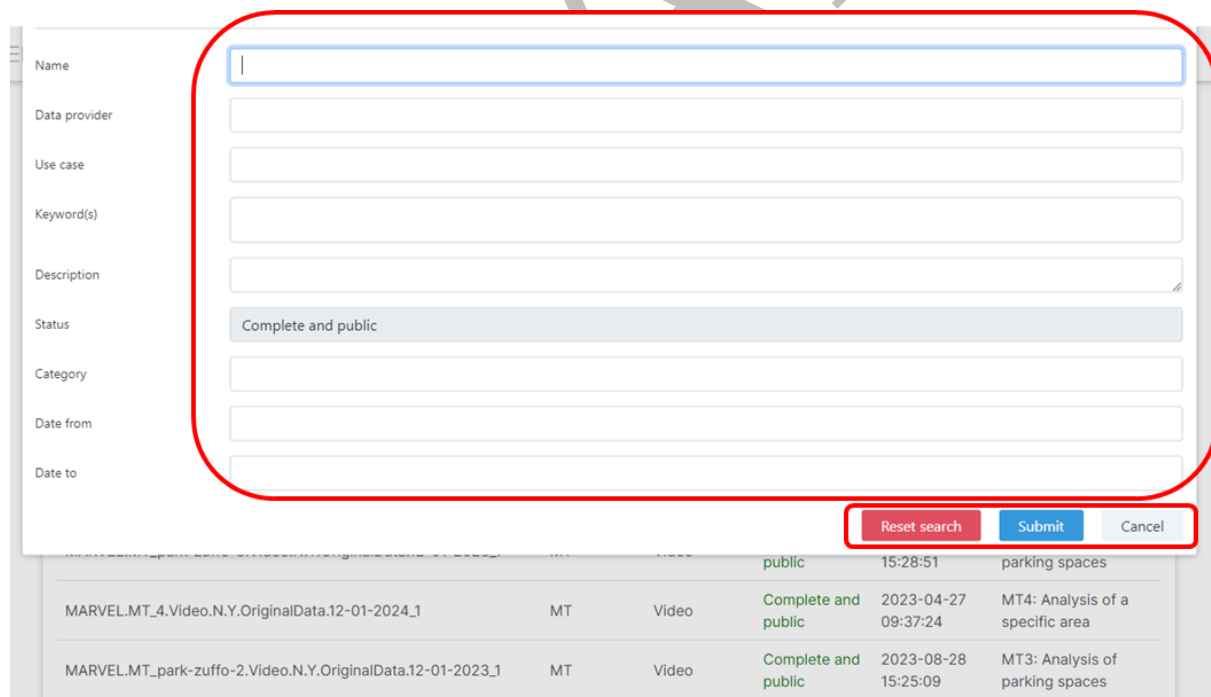


Figure 13. GUI – Datasets View 1-2 – Complex Search tab

4.3 Dataset detailed view

The detailed Dataset view shows all Dataset’s metadata (descriptive and technical) (Figure 14- Figure 15), along with the related list of Snippets (Figure 16). Next to each Snippet file (main

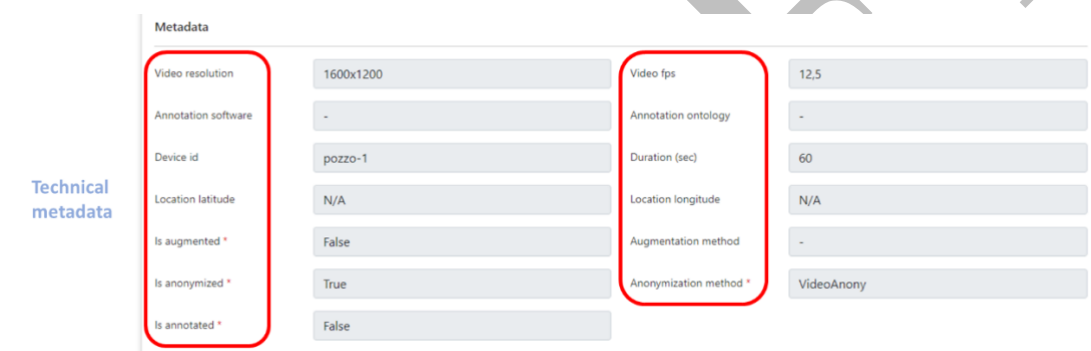
recorded file, annotations file, or inference results file), there is a download button. The user can navigate through this page and download all Dataset's files.



The screenshot shows the 'View dataset' page in the MARVEL interface. The 'Descriptive metadata' section is highlighted with a red box and includes the following fields:

- Name: MARVELMT_4.Video.N.Y.OriginalData.12-01-2024_1
- Data provider: MT
- Use case: MT4: Analysis of a specific area
- Keyword(s): Traffic trajectories, vehicle types, counting
- Description: Recording of the entrance to the pedestrian area in via Pozzo, near the train station
- Status: Complete and public
- Category: Video

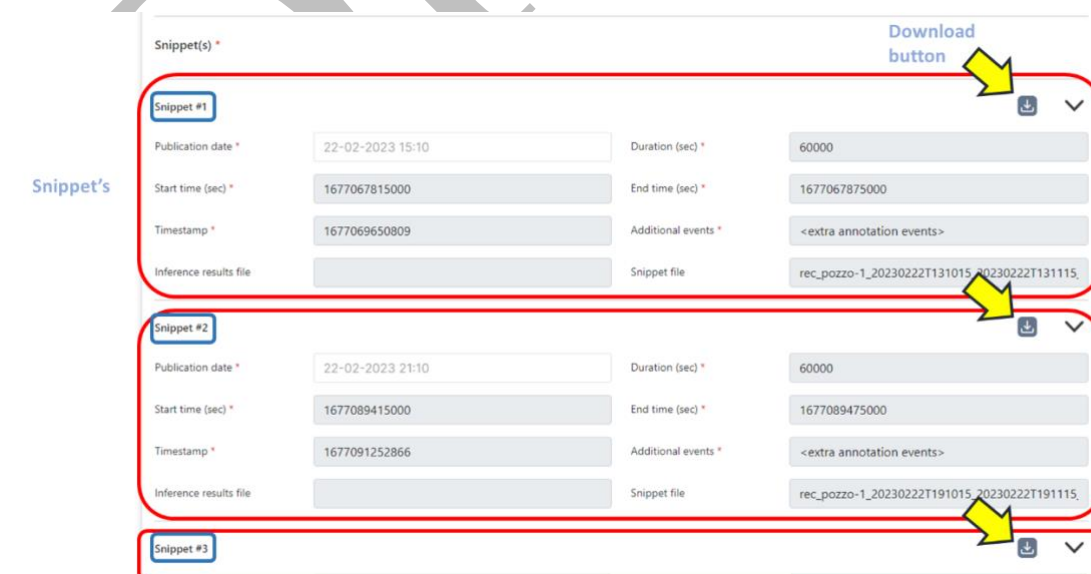
Figure 14. GUI – Dataset View 1-1 – Descriptive metadata



The screenshot shows the 'Metadata' section of the dataset view. The 'Technical metadata' section is highlighted with a red box and includes the following fields:

- Video resolution: 1600x1200
- Annotation software: -
- Device id: pozzo-1
- Location latitude: N/A
- Is augmented *: False
- Is anonymized *: True
- Is annotated *: False
- Video fps: 12.5
- Annotation ontology: -
- Duration (sec): 60
- Location longitude: N/A
- Augmentation method: -
- Anonymization method *: VideoAnony

Figure 15. GUI – Dataset View 1-2 – Technical metadata



The screenshot shows the 'Snippet(s)' list section of the dataset view. The list contains three snippets, each with a 'Download button' highlighted by a yellow arrow. The snippets are:

Snippet #	Publication date *	Duration (sec) *	Start time (sec) *	End time (sec) *	Timestamp *	Additional events *	Inference results file	Snippet file
Snippet #1	22-02-2023 15:10	60000	1677067815000	1677067875000	1677069650809	<extra annotation events>		rec_pozzo-1_20230222T131015_20230222T131115
Snippet #2	22-02-2023 21:10	60000	1677089415000	1677089475000	1677091252866	<extra annotation events>		rec_pozzo-1_20230222T191015_20230222T191115
Snippet #3								

Figure 16. GUI – Dataset View 1-3 – Snippets' list

4.4 Snippets view

Through the main page, the user can select the Snippets view tab (Figure 17). This is quite similar to the Datasets view tab, but for the Snippets. There are again a series of figures with aggregative information for all Snippets (i.e., Total number, Snippets per Category, and Snippets per Provider).

Once selecting a Snippet entry from the table, the detailed Snippet view is presented (see the next subsection 4.5).

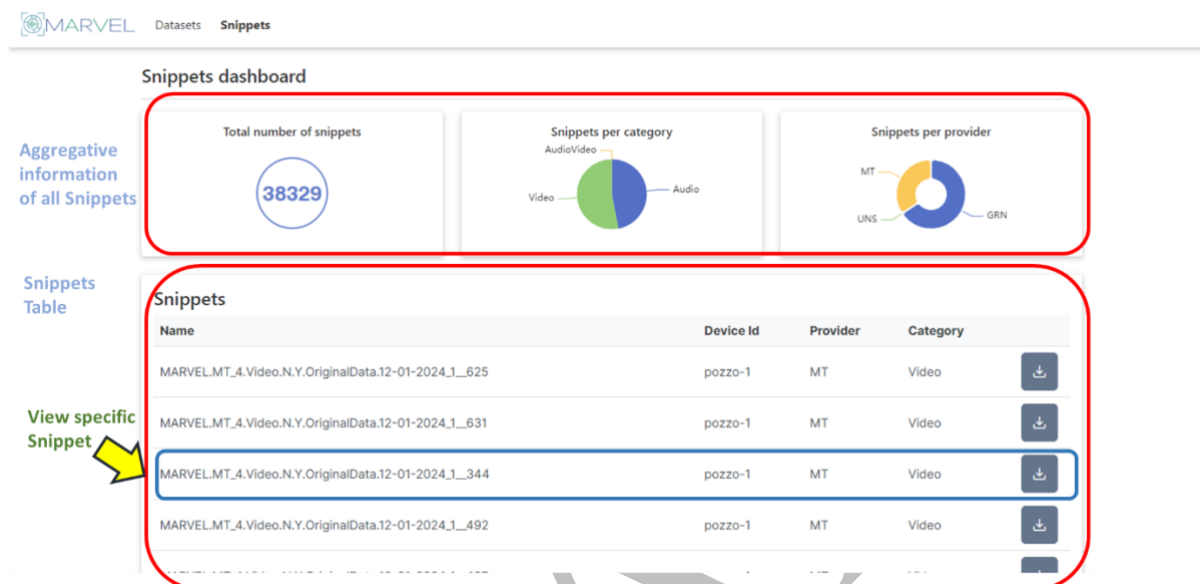


Figure 17. GUI – Snippets View – Snippets’ Aggregated data and Table

4.5 Snippet detailed view

This view is similar to the Dataset detailed view but for individual snippets (Figure 18). The user can review the Snippet details, like publication date, anonymisation/annotation/augmentation methods and captured events in the specific Snippet. There is also a download button to retrieve the data of the current snippet.

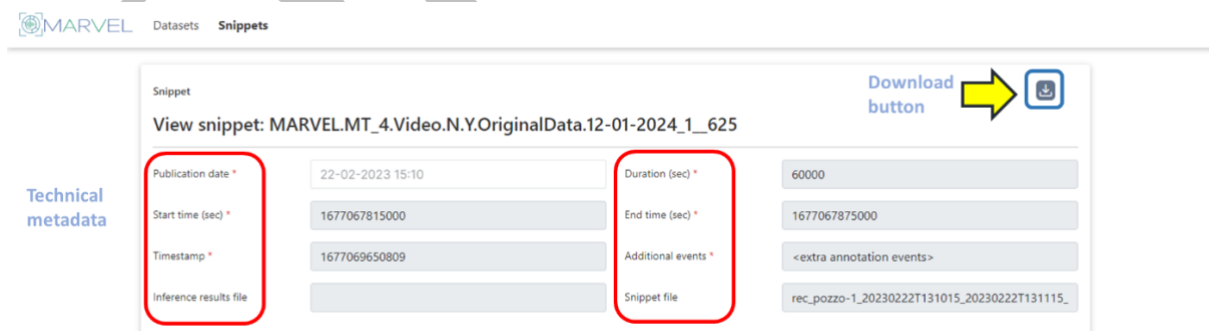


Figure 18. GUI – Snippet View – Technical metadata and Download button

4.6 MARVEL User Web Interface

The private web interface for the MARVEL Users is similar to the public one, offering additional functionality for *Insert*, *Update*, and *Delete*. This interface can be accessed under the unified *authentication and user management* mechanisms of the MARVEL Platform. Upon agreement, access can be given to external users, who want to contribute to our

community with their own datasets. The link is available at: <https://datacorpus.marvel-platform.eu/>.

4.7 Programmatic Interfaces & APIs

4.7.1 Public APIs

Apart from the graphic interfaces, there are also **APIs** that someone can use to **retrieve high volumes of data programmatically**. This subsection provides the technical details for the public Corpus APIs that someone can utilise to retrieve high volumes of data programmatically. These **public APIs** do not require user authentication and are mainly supporting the functionalities of **View**, **Search**, and **Download**.

The public Corpus GUI is available in the following link: $\$PUBLIC_URL = \text{https://datacorpus.marvel-project.eu/}$.

4.7.1.1 Public APIs for Datasets

Table 3. API – View all Datasets

Description	View summarised information about all available datasets
Link	$\$PUBLIC_URL/backend/info$
Input	No input is required
Output	JSON table

Table 4. API – Search Datasets

Description	Search on datasets
Link	$\$PUBLIC_URL/backend/marvel/v1/datasets-complex-search/$
Input	Optional values of the following: $?dataset$, Default Value = "", String dataset, $?dataProvider$, Default Value = "", String dataProvider, $?useCase$, Default Value = "", String useCase, $?keywords$, Default Value = "", String keywords, $?description$, Default Value = "", String description, $?category$, Default Value = "", String category, $?dateFrom$, Default Value = "", String dateFrom, $?dateTo$, Default Value = "", String dateTo,
Output	JSON table

Table 5. API – View specific Dataset

Description	View specific dataset based on the <i>key</i> field
Link	$\$PUBLIC_URL/backend/dataset/marvel/v1/datasets/{key}$

Input	Dataset's key (e.g., 0ff3c744-f903-4d20-a92a-eb62d7beaaad)
Output	JSON table

4.7.1.2 Public APIs for Snippets

Table 6. API – View all Snippets

Description	View summarised information about all available snippets
Link	$\$PUBLIC_URL/backend/snippets$
Input	No input is required
Output	JSON table

Table 7. API – View specific Snippet

Description	View all snippets of a dataset based on the <i>key</i> field
Link	$\$PUBLIC_URL/backend/marvel/v1/snippetsofdataset/{key}$
Input	Dataset's key (e.g., 56f47eed-acb8-4701-934a-744252a835fc)
Output	JSON table

Table 8. API – Download specific Snippet

Description	Download specific snippet file based on the <i>file name</i>
Link	$\$PUBLIC_URL/backend/marvel/v1/downloadsnippet/{key}$
Input	Snippet's key (e.g., c6bdce81-5660-4ec4-ac60-35622e48d9f0)
Output	File

4.7.2 Private APIs

There are also *private APIs* that are utilised internally by other MARVEL components and the private Corpus GUI, offering functionality of *Insert*, *Update*, and *Delete*, as well as some pre-defined *queries* and *metrics* (e.g., total number of datasets and snippets, datasets/snippets per provider or dataset type, etc.). These APIs can be consumed under the unified *authentication and user management* mechanisms of the MARVEL Platform. Upon agreement with the technical team, a detailed API description can be provided for someone who wants to *use the Corpus service programmatically*. Moreover, *potential external contributors to our community*, who want to publicly share their data through our platform, are also welcome.

5 Augmentation Engine

5.1 Augmentations in ML

Another valuable aspect of the Data Corpus involves integrating *augmentation methods*. In the field of machine learning, it is common practice to modify the original datasets with the aim of improving the categorisation or other capabilities of the trained models. For instance, if someone has gathered a video dataset from a public square in the morning, they can generate an augmented variant of this dataset by applying brightness filters and simulating identical scenarios during the afternoon. Consequently, the improved machine learning model is anticipated to exhibit superior performance when assessing an authentic video stream in the afternoon, compared to the original model trained exclusively on the initial raw data.

Thus, data augmentations are employed in various machine learning tasks to enhance the performance and robustness of models. The primary reasons for using data augmentations include:

- **Increased Diversity:** Data augmentation introduces diversity into the training dataset by applying various transformations to the existing data. This helps the model generalise better to unseen examples and different conditions.
- **Improved Generalisation:** By presenting the model with a wider range of variations in the training data (such as rotations, flips, zooms, etc.), it learns to recognise patterns more robustly and is less likely to overfit to specific instances in the training set.
- **Addressing Limited Data:** In situations where the available labelled data is limited, data augmentation can artificially increase the size of the training dataset, providing the model with more examples to learn from.
- **Enhanced Invariance:** Augmentations help in building invariance to certain transformations. For example, if a model is trained with augmented images that include various rotations, it is likely to become rotationally invariant.
- **Regularisation:** Data augmentations act as a form of regularisation, helping to prevent overfitting by discouraging the model from relying too heavily on specific features present in the training data.
- **Realism and Variability:** Augmented data can represent a more realistic and variable set of scenarios. For example, in computer vision tasks, images with different lighting conditions, orientations, and backgrounds can improve the model's performance in real-world situations.

Overall, data augmentations contribute to creating more robust and adaptable machine learning models, particularly in scenarios where the available labelled data is limited or when the model needs to perform well under diverse conditions.

5.2 Implementation and Supported Augmenters

For the goals of Data Corpus, an *Augmentation Engine* was deployed within one of its VMs. This engine supports several augmentation techniques, both for video and audio files, trying also to simulate different timepoints within the day or different weather conditions (e.g., apply a video filter to simulate rain). To facilitate ML experts, a continuous augmentation of the ingested datasets can be activated. An augmentation strategy has been developed, including state-of-the-art techniques in terms of augmenting AV data. Based on this strategy, several Python-based augmentation scripts were deployed and tested to automatically augment selected

datasets. Augmentation examples along with the use of these scripts have been presented in *D2.2 – Management and distribution Toolkit – initial version* [18].

On-demand, the user could choose the augmentation function along with its parameters via a Command Line Interface (CLI), based on his/her needs. The script further takes as arguments the input folder where the original non-augmented files reside and the output folder where the augmented files will be produced. The script also takes care of the name production of the augmented files while it prints for each augmentation the respective augmentation parameters.

The open-source software libraries of Keras [6], TensorFlow [7], imagaug [8], audiomentations [9], and torch-audiomentations [10], can be used for augmentations within the Corpus deployment. Table 9 presents the most indicative augmenters that are supported.

Table 9. Video and Audio Augmenters

Augmenter	Description
<i>Video/Image</i>	
Arithmetic	Add a value to all pixels in a frame
Artistic	Convert the style of frames to a more cartoonish one
Blend	Alpha-blend two image/frame sources using an alpha/opacity value
Blur	Augmenter to blur frames using Gaussian kernels
Collections	Apply RandAugment to inputs as described in the corresponding paper
Colour	Apply child augmenters within a specific colorspace
Contrast	Adjust frame contrast by scaling pixel values
Convolutional	Apply a Convolution to input frames
Debug	Visualise data in batches and save corresponding plots to a folder
Edges	Apply a canny edge detector to input frames
Flip	Horizontal, vertical, or mirror flip of frames
Geometric	Augmenter to apply affine transformations to images
Imgcorruptlike	Add Noise (Gaussian, Shot, Impulse, Speckle), Blur (Gaussian, Glass, Defocus, Motion, Zoom), Fog, Frost, Snow, Spatter, Contrast, Brightness, Saturate, Jpeg Compression, Pixelate, or Elastic Transformation
Pillike	Augmenters that have identical outputs to well-known Python Imaging Library (PIL) functions (Solarise, Posterise, Equalise, Autocontrast, EnhanceColor, EnhanceContrast, EnhanceBrightness, EnhanceSharpness, FilterBlur, FilterSmooth, FilterSmoothMore, FilterEdgeEnhance, FilterEdgeEnhanceMore, FilterFindEdges, FilterCountour, FilterEmboss, FilterSharpen, FilterDetail, and Affine)
Pooling	Apply average pooling to frames
Segmentation	Completely or partially transform images to their superpixel representation
Size	Augmenter that resizes images to specified heights and widths

Augmenter	Description
Weather	Augmenters that create weather effects (FastSnowyLandscape, CloudLayer, Clouds, Fog, SnowflakesLayer, Snowflakes, RainLayer, Rain)
Meta	List augmenter that may contain other augmenters to apply in sequence or random order
Audio	
AddBackgroundNoise	Mixes in another sound to add background noise
AddGaussianNoise	Adds Gaussian noise to the audio samples
AddGaussianSNR	Injects Gaussian noise using a randomly chosen signal-to-noise ratio
AddShortNoises	Mixes in various short noise sounds
AddColoredNoise	Add coloured noise to the input audio
AdjustDuration	Trims or pads the audio to fit a target duration
AirAbsorption	Applies frequency-dependent attenuation simulating air absorption
ApplyImpulseResponse	Convolve the audio with a randomly chosen impulse response
BandPassFilter	Applies band-pass filtering within randomised parameters
BandStopFilter	Applies band-stop (notch) filtering within randomised parameters
Clip	Clips audio samples to specified minimum and maximum values
ClippingDistortion	Distorts the signal by clipping a random percentage of samples
Gain	Multiplies the audio by a random gain factor
GainTransition	Gradually changes the gain over a random time span
HighPassFilter	Applies high-pass filtering within randomised parameters
HighShelfFilter	Applies a high shelf filter with randomised parameters
Identity	This transform returns the input unchanged and it can be used for simplifying the code in cases where data augmentation should be disabled
Lambda	Applies a user-defined transform
Limiter	Applies dynamic range compression limiting the audio signal
LoudnessNormalization	Applies gain to match a target loudness
LowPassFilter	Applies low-pass filtering within randomised parameters
LowShelfFilter	Applies a low shelf filter with randomised parameters
MP3Compression	Compresses the audio to lower the quality
Normalize	Applies gain so that the highest signal level becomes 0 decibels relative to full scale (dBFS)
Padding	Replaces a random part of the beginning or end with padding
PeakingFilter	Applies a peaking filter with randomised parameters
PeakNormalization	Apply a constant amount of gain, so that highest signal level present in each audio snippet in the batch becomes 0 dBFS

Augmenter	Description
PitchShift	Shifts the pitch up or down without changing the tempo
PolarityInversion	Flips the audio samples upside down, reversing their polarity
RepeatPart	Repeats a subsection of the audio a number of times
Resample	Resamples the signal to a randomly chosen sampling rate
Reverse	Reverses the audio along its time axis
RoomSimulator	Simulates the effect of a room on an audio source
SevenBandParametricEQ	Adjusts the volume of 7 frequency bands
Shift	Shifts the samples forwards or backwards
ShuffleChannels	Given multichannel audio input (e.g., stereo), shuffle the channels, e.g., so left can become right and vice versa
SpecChannelShuffle	Shuffles channels in the spectrogram
SpecFrequencyMask	Applies a frequency mask to the spectrogram
TanhDistortion	Applies tanh distortion to distort the signal
TimeMask	Makes a random part of the audio silent
TimeStretch	Changes the speed without changing the pitch
TimeInversion	Reverse (invert) the audio along the time axis similar to random flip of an image in the visual domain
Trim	Trims leading and trailing silence from the audio

For scalability and performance, the various augmentations can be executed in parallel, while some of the augmenters support internal parallelisation and exploitation of multi-core environments, both with CPUs and GPUs.

After consolidating with ML experts and practitioners, two *augmentation profiles* were created for video and audio augmentations, respectively. In these profiles, a set of pre-defined augmenters have been selected, and the user can activate this and apply it in selected datasets in a semi-automated fashion. These profiles include:

- Video augmentations profile
 - Gaussian noise, Severity level, $1 \leq \text{severity} \leq 5$
 - Brightness, Severity level, $1 \leq \text{severity} \leq 5$, imitate night-time conditions (lower brightness)
 - Grayscale, value =1 (mostly the new grayscale image is visible)
 - Sharpen
 - Linear Contrast, Severity level, $1 \leq \text{severity} \leq 5$
 - Saturate, Severity level, $1 \leq \text{severity} \leq 5$
 - JpegCompression, Severity level, $1 \leq \text{severity} \leq 5$
 - Pixelate, Severity level, $1 \leq \text{severity} \leq 5$
 - Horizontal flip

- Random rotation
- Random crop
- Audio augmentations profile
 - Pitch scaling
 - Time stretching
 - Add Gaussian noise with random signal-to-noise ratio (SNR)
 - Add Short Noises to simulate burst-like noise
 - Air Absorption to simulate microphone distance to sound sources
 - Limiter to simulate automatic gain control and dynamic range compression
 - MP3 Compression
 - Resample audio into a lower sample rate

For the Gozo streams, the video augmentation profile has been fully deployed and automated. After the anonymisation of the video streams by VideoAnony, the files are transmitted to the Corpus infrastructure for further processing. There is one folder for each one of the 7 Gozo streams, where these files will be stored accordingly.

A process reads the contents of these folders. For each original file, the video augmentation profile is executed. The produced augmentation files are stored in a folder with the name of the original file. When all augmentations for a file are completed, the process compresses the related folder.

In parallel, two ingestion flows are deployed with the ingestion tool. The first one creates a new dataset entry and uploads only the original files. The second one creates an augmented dataset entry and uploads the compressed file as snippets. Figure 19 depicts the result as it can be viewed in the public GUI, after filtering with the dataset's name.

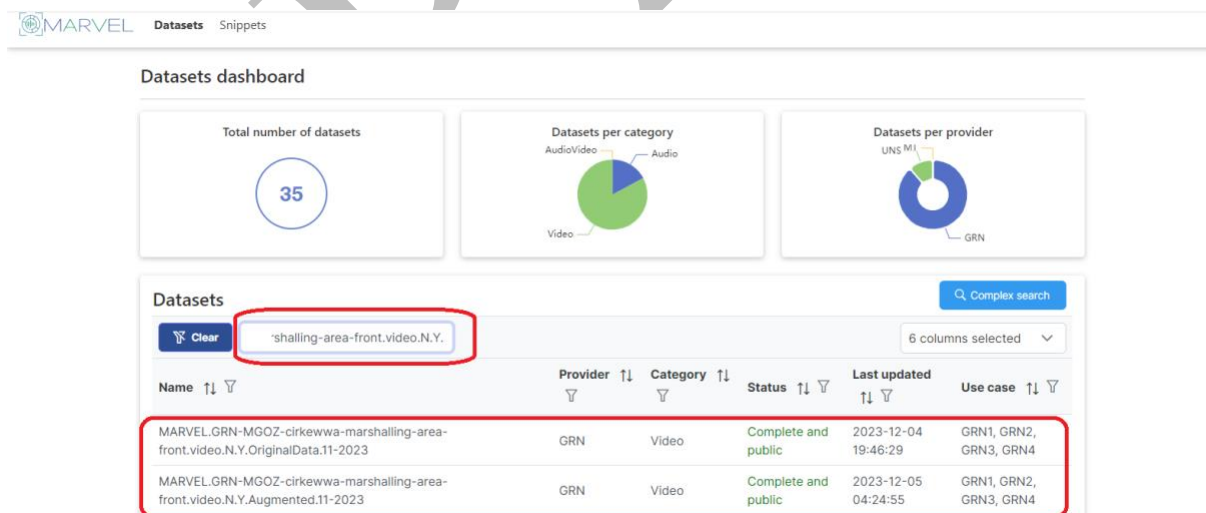


Figure 19. Original and Augmented datasets for one Gozo stream

6 The Ingested Datasets

This Section presents: *i)* the use cases, *ii)* stream sources, and *iii)* the datasets (original and augmented) that are available in the private and public interfaces of the Corpus.

6.1 Use cases

Datasets have been produced from all *10 piloting use cases* [2] of:



- MT
 - MT1: Monitoring of Crowded Areas
 - MT2: Detecting Criminal and Anti-Social Behaviours
 - MT3: Monitoring of Parking Places
 - MT4: Analysis of a Specific Area
- GRN
 - GRN1: Safer Roads
 - GRN2: Road User Behaviour
 - GRN3: Traffic Conditions and Anomalous Events
 - GRN4: Junction Traffic Trajectory Collection
- UNS
 - UNS1: Drone Experiment
 - UNS2: Audio-Visual Emotion Recognition






No specific use case is defined for the smart city of Gozo, but the processing goal is to enhance urban planning and traffic management, as for GRN.






6.2 Stream sources

In total, **29 distinct stream sources** have been exploited. From MT, 17 surveillance cameras have been integrated from the areas Duomo Garibaldi, Duomo Nord, Piazza Fiera, Piazza SMMaggiore, Piazza Dante, Park Zuffo, Zuffo Berlino, Zuffo SS Cosma-Damiano, Dogana, and Pozzo. GRN utilised 3 surveillance cameras around the city of Malta. UNS run 2 main use cases with drone recordings. From Gozo 7 public cameras have been provided, producing different datasets for each camera. Video recordings have been captured of the vehicle queuing area at the Gozo Channel ferry terminal, as well as from the marshalling front area, the Mgarr shore street (lower, middle, and upper street views), and the Cirkewwa Marshalling area (front and side views). Augmentations for these datasets were also produced in a semi-automated fashion. In general, the cameras could capture both video and audio data, but in most scenarios, only one of the options was used, based on the use cases' minimum requirements. Table 10 presents the stream sources that were used by MARVEL.

Table 10. Stream sources

Stream source	Data type	Use case	Camera view
MT			
Duomo-Garibaldi	Video	MT1	
Duomo-Nord	Video	MT1	
Sud	Video	MT1	
Piazza-Fiera-1	Video	MT1	
Piazza-Fiera-3	Video	MT1	
Piazza-Fiera-4	Video	MT1	
Piazza-SMMaggiore-TrafficLight	Audio	MT2	-
Piazza-SMMaggiore-Obelisque	Video	MT2	
Park-Zuffo-2	Video	MT3	
Park-Zuffo-3	Video	MT3	
Zuffo-Berlino	Audio	MT3	-
Zuffo-SS-Cosma-Damiano	Audio	MT3	-
Dogana-2	Video	MT4	
Dogana-3	Video	MT4	
Piazza-Dante-Listone	Video	MT4	
Piazza-Dante-TrafficLight	Video and Audio-Video	MT4	
Pozzo-1	Video	MT4	
GRN			
GRN-stream1	Video and Audio	GRN1-GRN4	
GRN-stream2	Audio and Audio-video	GRN1-GRN4	

Stream source	Data type	Use case	Camera view
GRN-stream3	Audio and Audio-video	GRN1-GRN4	
UNS			
Drone1	Video	UNS1	
Drone2	Audio	UNS2	
Gozo			
Cirkewwa-marshalling-area-front	Video	GRN1-GRN4	
Cirkewwa-marshalling-area-side	Video	GRN1-GRN4	

Stream source	Data type	Use case	Camera view
Mgarr-marshalling-area-front	Video	GRN1-GRN4	
Mgarr-road	Video	GRN1-GRN4	
Mgarr-shore-street-lower	Video	GRN1-GRN4	
Shore-street-middle	Video	GRN1-GRN4	
Shore-street-upper	Video	GRN1-GRN4	

6.3 Datasets

Each of the constructed datasets contains snippet files from a single streaming source. Currently, there are around 52,000 snippet files (audio, video, and audio-video) within around 64 datasets, 12 of which are augmented versions of the original ones. From them, 35 datasets and 28,500 snippets from GRN, UNS, and Gozo are public (the 29 datasets of MT are confidential for the moment until the privacy authorities of Italy come to a final decision

concerning the dissemination level of this data). The *total size* of the Corpus until M36 is around *1.IPBs*. Figure 20 shows the overall snippets' view in the private GUI.

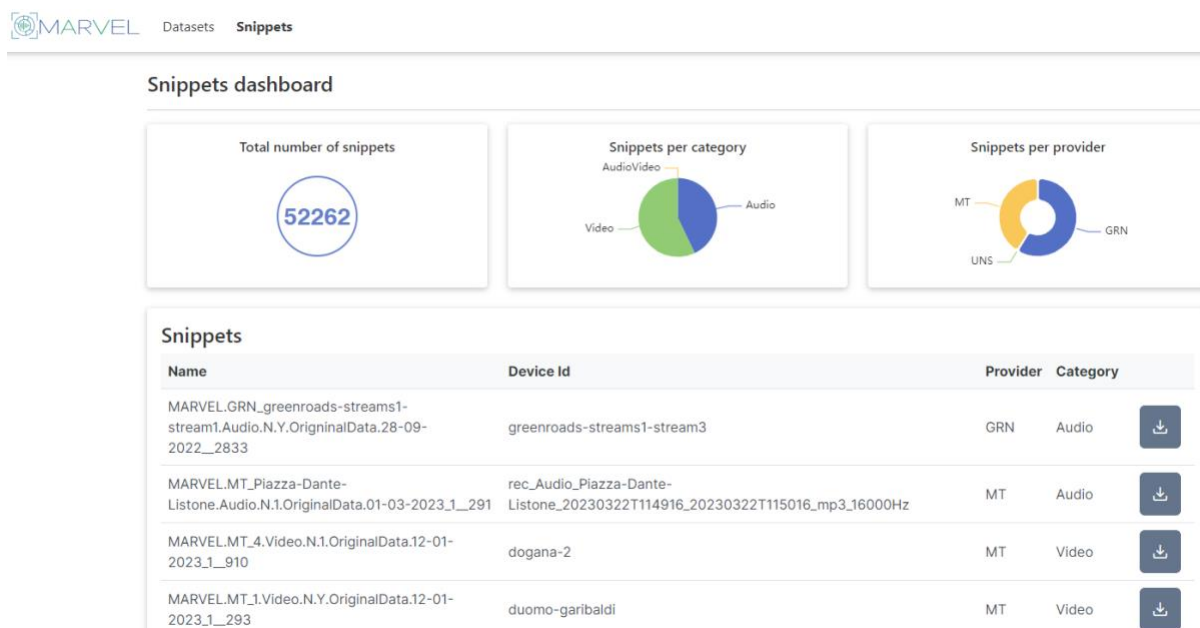


Figure 20. Total snippets volume

7 Evaluation Studies

This Section presents the overall evaluation activities that were conducted for the Data Corpus. These include assessments for the: i) performance, ii) user-friendliness, iii) security, and iv) privacy aspects.

7.1 Performance/Benchmarking

Throughout the project, several performance assessments were performed, including the main benchmarking trials that are documented under the deliverables *D5.2 – Technical evaluation and progress against benchmarks – initial version* [32] (M14) and *D5.5 – Technical evaluation and progress against benchmarks* [33] (M30).

In short, the Corpus fulfilled the performance-related requirements and goals. These mainly include: i) time constraints concerning storage, database queries, and end-user experience; ii) capability to process high volumes of data and many streaming sources; iii) scalability to the PBs order; and iv) correct operation of the database and the repository as a whole.

Table 11 summarises the benchmarking results, as reported in the latest D5.5. Data Corpus was operating as a distributed dockerised deployment in the PSNC infrastructure. The tested deployment had five datanodes, one namenode, and one region server. Thereupon, 10 ingestion requests were performed from parallel steams in several iterations, calculating the benchmarking metrics. At this stage, Data Corpus performed as expected.

Table 11. Performance benchmarking

Metric	Measurement
Data loss rate	None
Service availability – Failed request	None
Data access restriction	Accessing the HDFS with no access restriction in place (http)
Data transfer latency	~15ms (on average)
Data throughput	~120MB/s (on average)
Response time	~0.7ms (on average)

7.2 Ease-of-use & User Satisfaction

User-friendliness was an important factor for the user acceptance of the MARVEL Corpus solution. Three evaluation studies were conducted in total: two internal by consortium members and one by external users. The internal evaluations took place first. Several comments were made, improving the overall user experience. Once an acceptable level of user satisfaction was achieved, the external evaluation was held, confirming the good result.

The assessments were made in the form of Google Forms, where users of the Corpus filled out after some demonstration event, providing their feedback. The detailed responses are referred under the **Annex 3 – User Satisfaction Studies**.

7.2.1 Evaluation Study 1 – Internal

The first evaluation round was of great importance as it revealed from an early stage, several issues that could improve the user experience. The Data Corpus was used by 17 persons, who

then filled an extensive questionnaire, providing their feedback and fruitful comments. This included questions concerning the performance (e.g., time to search or download snippets and datasets), the usability (e.g., wherever the provided descriptions for datasets and their snippets were appropriate, if the search and filtering functionality was sufficient, etc.), and user-friendliness (e.g., how they rate the look-and-feel of the GUI). Several bugs were also discovered during this evaluation, which were later fixed. Moreover, the responders provided their opinion regarding the overall offering of the Corpus. All of them had a positive view and expressed the potential to make use of the Data Corpus for their activities or communicate it to their colleagues or other partners.

7.2.2 Evaluation Study 2 – Internal

After the initial evaluation study and its updates, the Corpus implementation continued, adding more features, and gradually improving further the user-friendliness aspects. The second evaluation round was completed by 5 persons, who used the final version of the Corpus and verified that it had reached a good level of user acceptance and could be further shared with external users from then. All responders mentioned good overall results and no negative aspects were recorded. Again, the responders were positive about the overall offerings of the Corpus and seemed willing to use it and/or promote it to their colleagues and partners or within their organisation.

7.2.3 Evaluation Study 3 – External

The third evaluation round was conducted by external users. Several experts had been contacted during the last phase of the project and were informed of the Data Corpus elements. A questionnaire was disseminated after related Info Day events, where these users could make use of the Corpus, and then, provide their feedback as well as their specific intention to use the various Corpus offerings.

Around 12 persons completed this activity. All of them provided very positive feedback and mentioned potential opportunities for collaboration. The main collaborations included: i) the use of data within ML courses or students' research/thesis; ii) utilise the data to train ML models that will be used in other activities or projects; iii) use the current MARVEL Data Corpus deployment to upload their own datasets; and iv) use the open-source implementation as a building block of their own Big Data solutions. Thereupon, a community of external stakeholders was formed consisting of academia and industrial groups, which could exploit the Data Corpus within the next period.

7.3 Security Assessments

The main security and vulnerability assessments of the Corpus Big Data infrastructure were performed by the STS's Assurance Platform [11]-[13]. The following subsections describe the Assurance Platform and the security assessments that were performed.

7.3.1 Assurance Platform description

STS's Security and Privacy suite takes on the responsibility of continuously monitoring, testing, and evaluating the security (and, if required, privacy) status of the protected organisation(s) and their assets in real-time. It incorporates various built-in security assessments that focus on the principles of Confidentiality, Integrity, and Availability (CIA), utilising custom metrics that can be adjusted to fit the components of the platform. This approach employs evidence-based methods, offering security assurance assessments with verifiable outcomes. The internal architecture of the Assurance Platform is outlined in a high-level perspective in Figure 21.

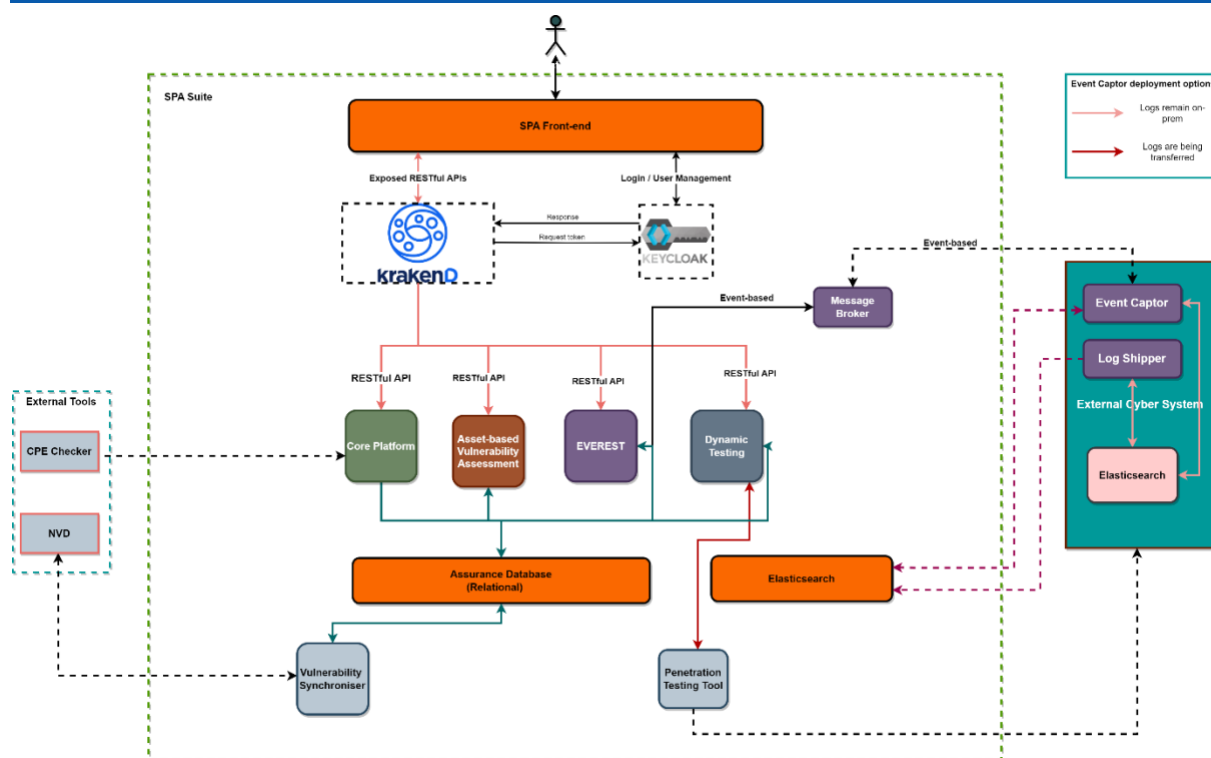


Figure 21. Assurance Platform – Internal architecture

7.3.1.1 Core Platform

The component holds the responsibility of overseeing the asset and assessment model for an organisation. It consists of two integral parts: the Asset Loader and the Assessment Loader. The former maintains the asset model for the organisation's cyber system, encompassing assets, their associated security properties, relationships between assets, and the security controls safeguarding them. Users can input assets using the SPA Suites' front end through wizards, by completing a pre-structured Excel sheet and uploading it via the front end or the API, or by submitting a Software Bill of Materials (SBOM) in CycloneDX or SPDX format. The Assessment Loader, on the other hand, manages the available assessments within the SPA Suite, handling criteria, profiles, results, and model executions.

7.3.1.2 Asset-based vulnerability assessment

This component is tasked with conducting a passive vulnerability assessment to detect known vulnerabilities associated with assets specified in an organisation's asset model. It interacts with the core platform, obtaining the Common Platform Enumeration (CPE) [35] of assets during the assessment execution. Subsequently, it retrieves pertinent Common Vulnerabilities and Exposures (CVE) entries [14] by searching a local version of the National Vulnerability Database (NVD) [36]. This database, maintained by the National Institute of Standards and Technology, is regularly updated through an in-house component that fetches the latest CVEs. Following the vulnerability assessment, the tool generates a report on known vulnerabilities for assets possessing a valid CPE.

7.3.1.3 Dynamic Tester

This module is tasked with instigating dynamic testing assessments, encompassing activities like penetration testing or active vulnerability assessments, targeted at a specific organisation. The execution of these assessments necessitates an agreed-upon and signed Target of Evaluation (ToE) document from the organisation. The dynamic tester within this module also

has the capability to identify assets not incorporated in the existing asset model. Its responsibilities extend to parsing the report generated by the corresponding tool, generating a new assessment model execution inclusive of the corresponding results, and incorporating the newly identified assets into the organisation's asset model. Alternatively, upon successful completion of the assessment, a comprehensive report is dispatched to the organisation, conveying pertinent information extracted from the evaluation.

7.3.1.4 Event REaSoning Toolkit (EVEREST)

The module tasked with overseeing the target organisation's cyber system for potential issues operates under the moniker EVEREST. In its monitoring capacity, EVEREST adopts a comprehensive approach, scrutinising critical facets of the cyber system, including network traffic, potential threats from both internal and external sources, misconfigurations, and the correct installation of security controls. Utilising a continuous evaluation mechanism against predefined rules articulated in Event Calculus and Drools, EVEREST excels in detecting anomalies, deviations, and potential risks within the system. Collaborating with Event Captors, EVEREST retrieves raw events directly from the cyber system.

7.3.1.5 Event Captors

An Event Captor functions as a tool that, adhering to specifications established by EVEREST, consolidates log and event data from the targeted infrastructure and encapsulates it into a specific format compatible with the EVEREST model. There are two modes for collecting logs and events. The first mode is grounded in the ELK solution, where Elasticsearch and lightweight shippers like Beat are employed to transmit and centralise log data. The second mode involves STS's Native Event Captors, specifically designed captors that don't rely on the logging capabilities of the ELK stack. EVEREST initiates the necessary Event Captors as per its requirements.

7.3.1.6 Security Component

The component responsible for delivering Identity and Access Management functionalities, along with an API Gateway for the exposure of RESTful APIs when necessary, consists of two primary elements: Keycloak, utilised for identity and access management, and KrakenD, the API Gateway that collaborates with Keycloak for authorisation. Keycloak is an open-source platform designed for identity and access management, offering a centralised access point for modern applications, APIs, and microservices. Its capabilities include single sign-on, identity brokering, social login, robust user management, and authentication features. On the other hand, KrakenD serves as a lightweight, high-performance API Gateway focused on exposing both internal and external microservices to the external environment. It simplifies the complexity and routing logic of core services, enabling easy configuration and deployment of API Gateway setups through a straightforward, declarative configuration file. Additionally, KrakenD provides features like rate limiting, circuit breaking, and caching to effectively manage API performance and reliability.

7.3.2 Vulnerability assessments

Under the Vulnerability Assessment analysis, an Asset Model is created in the Assurance Platform, specifying the system components that materialise the Corpus (mostly the Hadoop and HBase counterparts) and will be the subject of the security investigation. The Platform maps the defined assets and their version with the CPE, and then, it can automatically disclose the applicable CVE records.

During the initial trial of this evaluation, it was discovered that in the early stages of the project, we had utilised a vulnerable version of Hadoop (v2.9). Several vulnerabilities had been identified that imposed a high risk. Table 12 summarises the most severe CVEs that were found, along with the risk calculated by the Common Vulnerability Scoring System (CVSS) [37] and the related Common Weakness Enumeration (CWE) records [38] that have been mapped by NVD. Most of these vulnerabilities have high-risk values, while three of them are considered very critical for security.

Table 12. Assurance Platform vulnerability assessment – Identified CVEs

CVE	Title	Description	CVSS
CVE-2021-25642	Apache Hadoop YARN remote code execution in ZKConfigurationStore of capacity scheduler	ZKConfigurationStore which is optionally used by CapacityScheduler of Apache Hadoop YARN deserializes data obtained from ZooKeeper without validation. An attacker having access to ZooKeeper can run arbitrary commands as YARN user by exploiting this.	8.8
<i>CWE-502</i>	<i>Deserialization of Untrusted Data</i>		
CVE-2022-25168	Command injection in org.apache.hadoop.fs.FileUtil.unTarUsingTar	Apache Hadoop's FileUtil.unTar(File, File) API does not escape the input file name before being passed to the shell. An attacker can inject arbitrary commands.	9.8
<i>CWE-78</i>	<i>Improper Neutralization of Special Elements used in an OS Command ('OS Command Injection')</i>		
CVE-2021-33036	Apache Hadoop Privilege escalation vulnerability	A user who can escalate to yarn user can possibly run arbitrary commands as root user.	8.8
<i>CWE-22</i>	<i>Improper Limitation of a Pathname to a Restricted Directory ('Path Traversal')</i>		
CVE-2021-37404	Heap buffer overflow in libhdfs native library	There is a potential heap buffer overflow in Apache Hadoop libhdfs native code. Opening a file path provided by user without validation may result in a denial of service or arbitrary code execution.	9.8
<i>CWE-787</i>	<i>Out-of-bounds Write</i>		
CVE-2022-26612	Arbitrary file write during untar on Windows	The unTar function uses unTarUsingJava function on Windows and the built-in tar utility on Unix and other OSes. As a result, a TAR entry may create a symlink under the expected extraction directory which points to an external directory. A subsequent TAR entry may extract an arbitrary file into the external directory using the symlink name. This however would be caught by the same targetDirPath check on Unix because of the getCanonicalPath call. However, on Windows, getCanonicalPath doesn't resolve symbolic links, which bypasses the check. unpackEntries during TAR extraction follows symbolic links which allows writing outside expected base directory on Windows.	9.8
<i>CWE-59</i>	<i>Improper Link Resolution Before File Access ('Link Following')</i>		

CVE-2020-9492	Apache Hadoop Potential privilege escalation	WebHDFS client might send SPNEGO authorisation header to remote URL without proper verification.	8.8
<i>CWE-863</i>	<i>Incorrect Authorisation</i>		
CVE-2018-11765	Potential information disclosure in Apache Hadoop Web interfaces	Any users can access some servlets without authentication when Kerberos authentication is enabled and SPNEGO through HTTP is not enabled.	7.5
<i>CWE-287</i>	<i>Improper Authentication</i>		
CVE-2018-11768	Apache Hadoop HDFS FSImage Corruption	The user/group information can be corrupted across storing in fsimage and reading back from fsimage.	7.5
<i>CWE-119</i>	<i>Improper Restriction of Operations within the Bounds of a Memory Buffer</i>		
CVE-2018-8029	Apache Hadoop Privilege escalation vulnerability	A user who can escalate to yarn user can possibly run arbitrary commands as root user.	8.8
<i>NVD-CWE-noinfo</i>	<i>Insufficient Information</i>		
CVE-2018-11767	Apache Hadoop KMS ACL regression	KMS blocking users or granting access to users Incorrectly, if the system uses non-default groups mapping mechanisms.	7.4
<i>CWE-269</i>	<i>Improper Privilege Management</i>		
CVE-2018-1296	Apache Hadoop HDFS Permissive listXAttr Authorisation	HDFS exposes extended attribute key/value pairs during listXAttrs, verifying only path-level search access to the directory rather than path-level read permission to the referent.	7.5
<i>CWE-200</i>	<i>Exposure of Sensitive Information to an Unauthorised Actor</i>		
CVE-2018-8009	Apache Hadoop distributed cache archive vulnerability	Hadoop is exploitable via the zip slip vulnerability in places that accept a zip file.	8.8
<i>CWE-22</i>	<i>Improper Limitation of a Pathname to a Restricted Directory ('Path Traversal')</i>		

Thus, we had to upgrade our system with a latest version of Hadoop 3.3.4 which has resolved these issues. In the second iteration of the vulnerability assessment, no CVEs of high risk were discovered.

7.3.3 Assurance Profiles and CIA Monitoring

After the deployment of a secure implementation, Assurance Profiles were installed to monitor CIA aspects of the Corpus. These profiles are modelled in the EVEREST component of the Assurance Platform, with Event Captors gathering related information from the running system. The Corpus administrator can access the Platform's GUI and review the whole operation. Figure 22 depicts an Assurance Profile that has been started in EVEREST concerning the availability principle of an HTTP service during a test. An Event Captor is automatically deployed, which periodically checks the service's status (e.g., HTTP error codes) and notifies EVEREST. Initially, the service was up and running, and at some time-point it was stopped.

The profile summarises this information in several graphs and tables. The successful evaluations of the principle are depicted with green (the service was found to be running appropriately), while the unsuccessful ones are depicted with red (the service was found unavailable).

Such profiles were not just deployed during the various benchmarking and testing activities of the project but are running during the normal operation of the Corpus as well.



Figure 22. Assurance Platform GUI – Availability monitoring

7.4 Privacy Assessments

The European Union (EU) funded project SENTINEL (GA-101021659) [15] develops a platform for the *self-assessment of GDPR compliance* for SMEs. There was a collaboration between the two projects under the T2.3/T2.4 activities, where the *SENTINEL Platform* was utilised to assess the privacy controls and procedures of the MARVEL Data Corpus service.

Moreover, the main GDPR compliance monitoring of the MARVEL framework and the stored data in the Data Corpus was examined under task *T2.6 – Ethics, privacy and data protection compliance* and the detailed results are documented in the deliverable *D2.6 – Ethics, privacy, and data protection compliance management – final version*. This is a confidential document (not publicly available).

The following subsections are describing the SENTINEL Platform and its privacy assessment, as well as the summary of T2.6 outcomes.

7.4.1 SENTINEL Platform description

SENTINEL offers a series of tools that can be used for self-assessments. Complying with GDPR supposes for organisations to demonstrate that Operational and Technical Measures (OTMs) implemented to meet data protection requirements are appropriate and effective. It is then a twofold challenge to:

1. Identify data protection requirements, and
2. Determine OTMs to meet them.

7.4.1.1. SENTINEL Platform methodology and components

Evaluating compliance with GDPR consists of verifying whether OTMs are implemented, appropriate, and effective. GDPR Compliance Self-Assessment (CSA) has been developed to allow users to perform such verification. Based on the ISO/IEC 33000 processes assessment method, GDPR CSA uses a process assessment model that organises data protection requirements into six data protection processes: i) Record, ii) Personal Data Lifecycle Management (PDLM), iii) Rights, iv) Consent, v) Data Protection Management (DPMAN), and vi) Breach. These processes allow to structure the collection of information related to OTMs implemented. Assessment of their appropriateness and effectiveness depends on the privacy risk level of the Processing Activity (PA). Figure 23 illustrates the assessment scope based on the related privacy risk level.

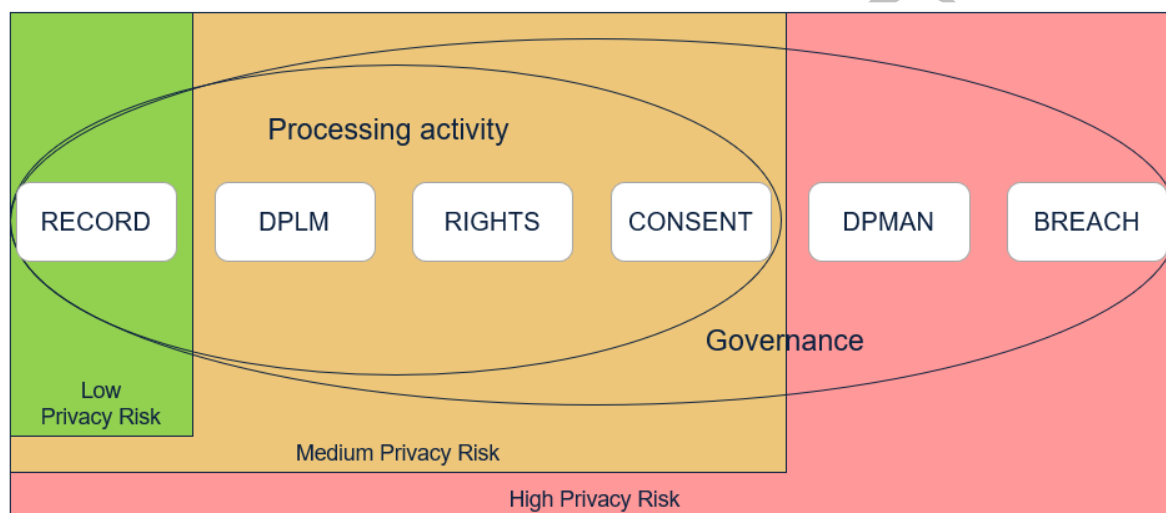


Figure 23. SENTINEL Platform – Assessment scope according Privacy Risk level

Similar to the Assurance Platform analysis above, the user starts the SENTINEL process by defining the system's assets in the Asset Inventory, as well as the PAs that will be analysed and those assets are participating in.

SENTINEL's Data Protection Impact Assessment (DPIA) toolkit was designed to allow organisations to identify (through assessment) and minimise (through recommendations) the risks associated with their personal data processing activities. The DPIA encompasses the following elements:

- A detailed description of the intended processing operations.
- The purposes behind the processing, including, when applicable, the legitimate interests pursued by the data controller.
- An evaluation of the necessity and proportionality of the processing operations concerning the stated purposes.
- An analysis of the potential risks to the rights and freedoms of data subjects.
- The anticipated measures designed to mitigate identified risks, incorporating safeguards, security protocols, and mechanisms ensuring the safeguarding of personal data.
- A demonstration of adherence to the GDPR, considering the rights and legitimate interests of data subjects and other relevant individuals.

Vulnerability assessment can also be performed for the participating assets, by a component called MITIGATE, similar to the Assurance Platform process. The effectiveness of the deployed security mechanisms and data protection controls is evaluated. SENTINEL's DPIA Toolkit is responsible for constructing the DPIA questionnaire and subsequently, for calculating the Processing Activities' risk based on the participant's responses. The CSA component aggregates all data and highlights the most critical problematic aspects along with a set of recommendations to cure them.

7.4.1.2. Privacy assessment

A PA was modelled in the SENTINEL Platform concerning the ingestion of data from the piloting environments to the MARVEL Data Corpus. The main assets that were included in the Asset Inventory are the latest version of Hadoop and HBase. Then, several questionnaires were filled along with the DPIA. Figure 24 depicts the MARVEL Corpus assessment profile under the SENTINEL Platform.

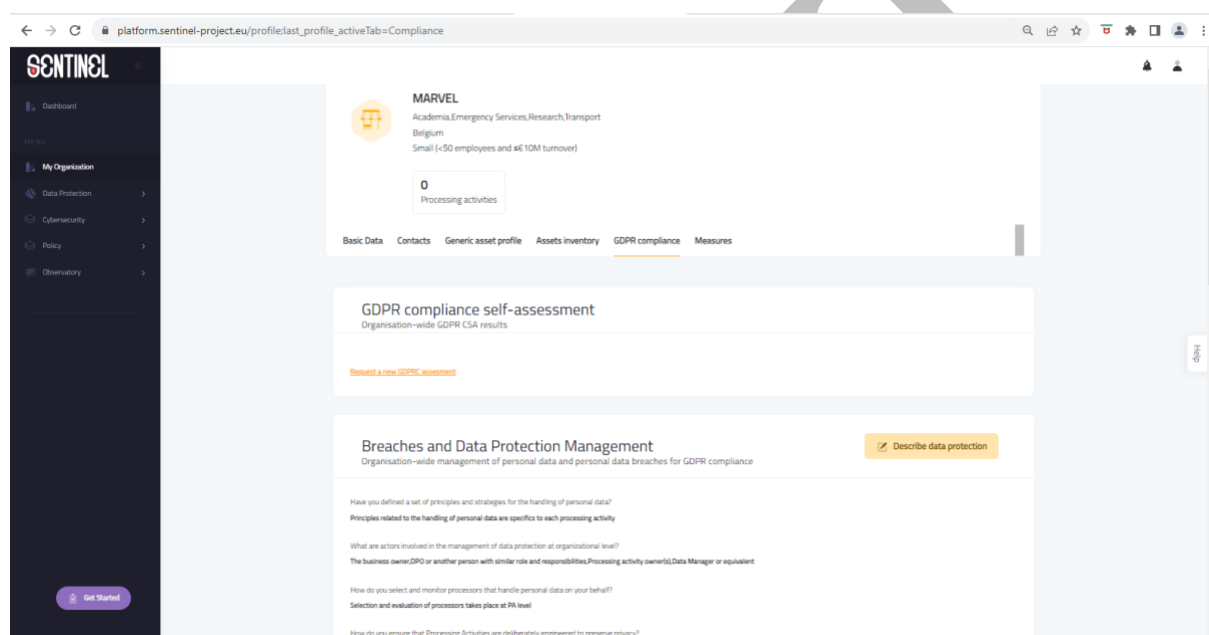


Figure 24. SENTINEL Platform – Privacy assessment

The main data protection was performed via the effective anonymisation of the AV streams. Data were protected in transit by cryptographic primitives, users were authenticated, and access control policies were enforced. However, the initial assessment brought to light deficiencies in the existing procedures for recording system activities and responding to GDPR-related requests from data owners. Consequently, significant enhancements were made to the monitoring and logging components (including the Assurance Platform and Zabbix agents), along with the refinement of procedures to effectively address requests from data owners, such as the project pilots, and other relevant entities.

The second assessment exhibited better results without any critical recommendation. However, it is stated that this is a self-assessment approach and not a formal DPIA. Nevertheless, this iterative process can ensure ongoing alignment with GDPR standards and reinforces the Corpus's commitment to privacy and data protection.

7.4.2 Ethics, privacy, and data protection compliance monitoring

As presented in D2.6, ethical, data protection, and privacy requirements have been properly identified and their transformation into actionable points has been successfully designed. Both

legal requirements and ethics requirements the project received, following a successful scientific evaluation, were properly addressed. They were submitted in the form of relevant deliverables. Pilots performed *internal DPIAs* before processing any data and sharing them with the Corpus. The MARVEL framework: i) follows the required *privacy-by-design* principles, ii) deploys the relevant *data privacy and security* elements, mainly anonymisation/pseudonymisation of data, E2F2C security, as well as other safeguards to protect data, iii) assigns *accountability* principles, and iv) uses appropriate *privacy governance* operations.

An *EAB* was also established, ensuring adherence to privacy laws and regulations, handling of anonymisation processes where applicable, handling of informed consent processes, processes associated with data security, secondary use of data, involvement of third countries in the project, and more.

DRAFT

8 Community Building and Actions Towards the Data Economy Vision of Smart Cities

The collected datasets and implemented technologies were proven to be very useful for end-users and helped in building a decent *community of academic and industrial organisations*, which provided very positive feedback and are willing to use it for their activities. A *community* of around 27 (13 external and 14 consortium members) academic and industrial groups have expressed the potential of using the Corpus offerings for: i) the public datasets, ii) the open-source implementation of the repository, and iii) the current MARVEL Data Corpus deployment. The following subsections are summarising the potential uses that were proposed by the various groups of experts.

ML concepts:

- Pattern matching
- Movement detection and tracking
- Traffic monitoring
- Traffic modelling
- Urban planning
- Air-quality monitoring
- Automotive vehicle development
- Public safety
- Personalised federated learning realisation for extreme-scale analytics
- Optimisation of ML algorithm
- Video and audio augmentations
- Sound localisation in scenes
- Speech detection
- Audio analytics and intelligent audio analysis
- Acquisition and analysis of big audio data
- AI inference and decision-making
- Environmental, bioeconomy, and climate change

8.1 Universities and Colleges

Several academic organisations were conducted during Info Days, conferences, or other dissemination events. In total, professors and lecturers from **8 universities** and **2 colleges** have explicitly expressed their interest in exploiting the Corpus results for academic purposes. These institutions are from 6 countries across Europe (Greece, Cyprus, UK, Denmark, Serbia, and Finland) with hundreds or thousands of students in their Computer Science departments. Three of the universities are consortium members who took part in the implementation of the AI solutions of MARVEL. The list of collaborators includes:

- Hellenic Mediterranean University (HMU) (Greece)

- Aristotle University of Thessaloniki (AUTH) (Greece)
- University of Piraeus (UNIPI) (Greece)
- Metropolitan College (AMC) (Greece)
- Mediterranean College (MEC) (Greece)
- European University Cyprus (EUC) (Cyprus)
- University of South Wales (UK)
- Aarhus University (AU) (Denmark)
- University of Novi Sad (UNS) (Serbia)
- Tampere University (TAU) (Finland)

All institutes include ML courses, while in some cases these courses are mandatory in the undergraduate degree programs. Advanced courses in postgraduate degrees and PhD research are also supported.

The core exploitable result is the **public datasets** of the Corpus. The main use case is their inclusion *within courses* for *Machine Learning*, *Pattern Recognition*, and *Artificial Intelligence*. Apart from the general concepts, other academic topics of interest include *sound localisation in scenes*, *movement detection and tracking*, *traffic monitoring*, *public safety*, *personalised federated learning realisation for extreme-scale analytics*, *optimisation of ML algorithm*, *video and audio augmentations*, and *acquisition and analysis of big audio data*. The Corpus data can be also utilised for research, student thesis and dissertations, workshops, and hands-on hackathons. The use of the **open-source implementation** as a first step to build their **own repositories** also gains some attention, mostly under more *engineering-focused projects* and *Big Data research*.

8.2 Research Institutes

Researchers were mostly conducted during conferences and workshops. In total, **3 research institutes**, which are also consortium members, have examined the potential to further exploit the Corpus outcomes. The institutes are located in 2 countries across Europe (Greece, and Italy) with strong expertise in the fields of ML and Big Data. The research collaborators are:

- Foundation for Research and Technology – Hellas (FORTH) (Greece)
- Italian National Research Council (CNR) (Italy)
- Fondazione Bruno Kessler (FBK) (Italy)

The Corpus offerings that can be utilised are the **public datasets** within research activities, the **MARVEL Corpus deployment for storing other open datasets**, and the use of the **open-source implementation as a building-block to build an institutional repository and engineer of Big Data infrastructures**. Research expertise of these groups is related to *implementation of AI/ML algorithms and Federated Learning mechanisms*, *video and audio anonymisation*, and *development and optimisation of Big Data solutions*.

Moreover, during the conferences HiPEAC 2023 and Data Week 2023, there were dedicated workshops and panel discussions among EU-funded Big Data projects. There, it was discussed potential collaborations between the three projects that were present at the event (MARVEL, EVEREST, and DAPHNE) and other projects that were represented under the auspices of the Big Data Value Association (BDVA). Among the collaboration steps that were examined, was

the exchange of experience, best practices, and data. The MARVEL *public datasets* could be utilised by the other projects for environmental research, such as *traffic modelling*, *air-quality monitoring*, and *automotive vehicle development*. Also, it was examined the potential to store public datasets of EVEREST and DAPHNE to the *MARVEL Corpus deployment*.

8.3 SMEs

SMEs were conducted during conferences and exhibitions where MARVEL was presented. In total, 8 SMEs are planning to utilise the Corpus offerings. The companies have premises in 4 countries (Greece, Germany, Malta, and Switzerland) with expertise in ML/AI, traffic analysis, as well as information technology and cybersecurity. The collaborating SMEs are:

- Techpro Academy (Greece)
- DRAXIS (Greece)
- UBITECH (Greece)
- Raven Cybersecurity (Greece)
- AEGIS IT Research (Germany)
- GreenRoads (Malta)
- ZELUS (Greece)
- Sphynx Technology Solutions (Switzerland)

The Techpro Academy is powered by Deloitte, Netcompany Intrasoft, ALTAIR, and DataScouting, and provides professional training through Bootcamps in Greece. Dozens of experts attend the training every year with skill-building for Front-end, Back-end, Data Science, and Junior Full-Stack classes. There are some first discussions with tutors to utilise the *public datasets* of the MARVEL Corpus within the *Data Science programme*.

DRAXIS focuses on developing real-life environmental Information and Communications Technology (ICT) solutions and providing specialised environmental consultation services and tools. Their solutions target the fields of air quality, weather forecasting, energy, waste management, circular economy, sustainable agriculture, and e-government. DRAXIS has extensive experience in implementing Environmental Impact Assessment studies and participates in several national and European projects. Their intent is to examine further the use of MARVEL's *public datasets* for research and innovation activities within *environmental*, *bioeconomy*, and *climate change projects*.

UBITECH is a leading, highly innovative software house, systems integrator and technology provider, established to provide leading-edge intelligent technical solutions and consulting services to businesses, organisations and governments to allow efficient and effective secure access and communication with various heterogeneous information resources and services. Their cutting-edge expertise covers among others the fields of Cloud computing, Big Data and analytics, factories of the future, cyber-physical systems and the Internet of Things, 5G technologies, block-chains, digital security, energy efficiency, and e/m-health. They will examine the use of the *public datasets* for the *development of extreme-scale analytics services*, while they will consider utilising the *open-source Corpus project to build their own Big Data solutions*.

Raven Cybersecurity is specialising in a wide range of cybersecurity services. It provides security assessments, creation of cybersecurity learning and training courses, and software development solutions. Raven empowers businesses of all sizes to safeguard their sensitive data

and avert cyber-attacks by providing tailored solutions that address each client's distinct needs and priorities. Their goal is to utilise the *open-source Corpus project* to build a Big Data repository for cyber-security focused datasets.

AEGIS IT Research develops cutting-edge software technology for a plethora of industry, business and public sectors. Their solutions include digital forensics and cyber-security, incident detection, and advanced visualisation. The company could make use of the *open-source Corpus project* to build a Big Data repository for the goals of a Security Operations Center (SOC). There could also be the potential for further collaboration in examining the current *MARVEL Data Corpus deployment* to build visualisation and forensics security solutions for Big Data infrastructures.

GreenRoads produces AI traffic video analytics for Smart Cities. Their goal is to empower urban planners with the insights they need to make informed decisions when optimising transport networks and urban infrastructure. GreenRoads is one of the MARVEL pilots that produced and shared several open datasets as well as AI components (e.g., CATFlow). They can still utilise all available *datasets* to further enhance their *AI solutions for urban planning and traffic monitoring*.

ZELUS offers data visualisation solutions which empower domain experts to discover patterns, behaviours and correlations of data items via an extensible, visualisation-based data exploration environment. This includes the SmartViz component that was developed under the MARVEL project activities. ZELUS will examine to further utilise the *public datasets* of the Corpus to improve its *AI inference and decision-making technologies*.

Sphynx Technology Solutions specialises in Information Technology (IT) solutions, ML and AutoML research, and cybersecurity. Within MARVEL, Sphynx led the activities of the Data Corpus and developed the Augmentation Engine, while the Assurance Platform was utilised for the main vulnerability and security assessments of the Corpus. The intention is to exploit the *Augmentation Engine* and enhance the *augmentation and sharing of Big AV Data services*, as well as advance its expertise and products for *monitoring and securing Big Data infrastructures*.

8.4 Industry

Several large companies and industries were contacted withing exhibitions and other communication events. In total, 6 companies are considering the exploitation of Data Corpus. These companies are located in five countries (Belgium, Luxembourg, Germany, Spain, and Poland). Their expertise is in the ICT domain, including Data Science and AI, analytics, automotive and mobility, Internet of Things, energy-efficiency and decarbonisation, cloud computing and High-Performance Computing (HPC), digital transformation, blockchains, cybersecurity.

The collaborating industries are:

- SWORD (Belgium)
- Intrasoft International (INTRA) (Luxembourg)
- Infineon (IFAG) (Germany)
- ATOS (Spain)
- audEERING GmbH (AUD) (Germany)
- Poznan Supercomputing and Networking Center (PSNC) (Poland)

The Corpus offerings that can be utilised are the *public datasets* within research activities (mostly for audio analysis), the *MARVEL Corpus deployment for storing other open datasets*, and the use of the *open-source implementation to build IaaS and PaaS solutions*. Relevant research expertise of these companies is related to *audio analytics and intelligent audio analysis, data acquisition, storage of data streams, and HPC*.

DRAFT

9 Conclusion

This deliverable is the *final outcome of the tasks T2.3 and T2.4*. It documents the overall activities that were performed concerning the *materialisation of the MARVEL Data Corpus*.

The Corpus development followed the *implementation phases* of the project (MVP (M18), 1st prototype (M18), 2nd prototype (M30)). The required infrastructure was provided by PSNC. The main repository was implemented with the Hadoop Distributed File System (HDFS) and the database HBase, while the graphical interface was implemented in Angular 2.0. The Corpus is also supporting several augmentation techniques for audio and video streams based on the libraries of Keras, TensorFlow, imagaug, audiomentations, and torch-audiomentations.

Several *integration flows* have been also deployed. All data are anonymised before ingestion by AudioAnony and VideoAnony. Direct ingest from these two components has been implemented. Also, the MARVEL AI components can process the audio/video streams and produce inference results concerning events or anomalies that they have identified. Therefore, the involved stream segments are stored locally as files by the StreamHandler, while the inference results are published in Kafka topics. Thereupon, ingestion of files from the StreamHandler end is also supported, along with the consumption of the inferences from a dedicated Kafka topic.

Moreover, many *evaluation studies* were conducted. Concerning performance, two benchmarking trials were held under WP5 activities. Corpus operation fulfils the defined functional requirements. The user acceptance was assessed under three studies, two internal by MARVEL partners and one by external users. The Corpus was found ease-to-use and quite user-friendly. The ethics and GDPR-compliance were examined in two iterations under the activities of task T2.5. Corpus follows the required privacy-by-design principles and implements the adequate security and data protection controls. Also, an Ethics Advisory Board (EAB) was established to verify the results. Additional security and privacy assessments were performed by the STS's Assurance Platform and the SENTINEL Platform, respectively. These analyses disclose several problematic elements that were identified and corrected.

Datasets from all pilots (MT, GRN, UNS) have been ingested in the Corpus. Moreover, datasets from the collaborative smart city of Gozo have been also stored. At the end of the project, the Corpus size reached the 1.1PBs. Although this is lower than the targeted goal (3.3PBs), the final result was very useful and a *community* of around 27 academic and industrial groups have expressed the potential of using the Corpus offerings.

A *joint exploitation plan* has been defined. The pilots' datasets are disseminated under a Creative Common license and ML experts and practitioners can utilise them for their activities. The implementation of FORTH and STS can be further provided under an Apache 2.0 license to further promote open science activities. The Corpus repository solution can be sold as Infrastructure-as-a-Service (IaaS) or Platform-as-a-Service (PaaS) by PSNC's products. Apart from the main offerings, services for technical support, consultancy, and training can be provided by FORTH and STS. Moreover, the security and privacy monitoring with the Assurance Platform can be also included as an additional service in the core IaaS or PaaS.

This deliverable and the developed Data Corpus contribute to the completion of the KPIs **KPI-O5-E1-1-KPI-O5-E4-1**. D2.5 along with D2.6 constitute the final results of WP2. The delivery of the two documents contributes to the fulfilment of the milestone **MS8 – Long-term sustainability and commercialisation**.

10 References

- [1] MARVEL Data Corpus public website, 2023-2024. Available at <https://datacorpus.marvel-project.eu/>. (Accessed 10th December 2023).
- [2] Dragana Bajovic, et al., “MARVEL: Multimodal Extreme Scale Data Analytics for Smart Cities Environments”, 4th International Balkan Conference on Communications and Networking (BalkanCom), Novi Sad, Serbia, 20-22 September 2021, pp. 1-5. (DOI: 10.1109/BalkanCom53780.2021.9593258).
- [3] Hadoop, Apache. Available at <https://hadoop.apache.org/>. (Accessed 10th December 2023).
- [4] HBase, Apache. Available at <https://hbase.apache.org/>. (Accessed 10th December 2023).
- [5] Toshifa Hussain, Anirudh Sanga, Shweta Mongia, “Big Data Hadoop Tools and Technologies: A Review”, International Conference on Advancements in Computing & Management (ICACM), SSRN, Jaipur, India, April 2019, pp. 574-578. (DOI: 10.2139/ssrn.3462554).
- [6] François Chollet, “Keras: Deep Learning for humans”. Available at <https://keras.io/>. (Accessed 10th December 2023).
- [7] Google, “TensorFlow: An Open Source Machine Learning Framework for Everyone”. Available at <https://www.tensorflow.org/>. (Accessed 10th December 2023).
- [8] Alexandser Jung, “imgaug: image augmentation in machine learning experiments”. Available at <https://imgaug.readthedocs.io/en/latest/>. (Accessed 10th December 2023).
- [9] Iver Jordal, “audiomentations: a Python library for audio data augmentation”. Available at <https://github.com/iver56/audiomentations>. (Accessed 10th December 2023).
- [10] Asteroid-team, “torch-audiomentations: Audio data augmentation in PyTorch”. Available at <https://github.com/asteroid-team/torch-audiomentations>. (Accessed 10th December 2023).
- [11] SPHYNX security and privacy assurance platform, SPHYNX. Available at <https://www.sphynx.ch/security-assurance-service/>. (Accessed 10th December 2023).
- [12] Hatzivasilis, G., Ioannidis, S., Kalogiannis, G., Chatzimpyrros, M., Spanoudakis, G., Prieto, G.J., Morgan, A.R., Lopez, M.J., Basile, C. and Ruiz, J.F., 2023, July. Continuous Security Assurance of Modern Supply-Chain Ecosystems with Application in Autonomous Driving: The FISHY approach for the secure autonomous driving domain. In 2023 IEEE International Conference on Cyber Security and Resilience (CSR) (pp. 464-469). IEEE. (DOI: 10.1109/CSR57506.2023.10224971).
- [13] Lakka, E., Hatzivasilis, G., Karagiannis, S., Alexopoulos, A., Athanatos, M., Ioannidis, S., Chatzimpyrros, M., Kalogiannis, G. and Spanoudakis, G., 2022, June. Incident Handling for Healthcare Organizations and Supply-Chains. In 2022 IEEE Symposium on Computers and Communications (ISCC) (pp. 1-7). IEEE. (DOI: 10.1109/ISCC55528.2022.9912965).
- [14] Common Vulnerabilities and Exposures (CVE), MITRE. Available at <https://cve.mitre.org/>. (Accessed 10th December 2023).

- [15] SENTINEL Project, Bridging the security, privacy and data protection gap for smaller enterprises in Europe, EU GA-101021659, 2021-2024. Available at <https://sentinel-project.eu/>. (Accessed 10th December 2023).
- [16] General Data Protection Regulation 2016. Available at <https://eur-lex.europa.eu/eli/reg/2016/679/oj>. (Accessed 10th December 2023).
- [17] “D2.1: Collection and Analysis of Experimental Data,” Project MARVEL, 2021. <https://doi.org/10.5281/zenodo.5052713>. (Accessed 10th December 2023).
- [18] “D2.2: Management and distribution Toolkit – initial version,” Project MARVEL, 2022. <https://doi.org/10.5281/zenodo.6821195>. (Accessed 10th December 2023).
- [19] “D2.4 - Management and distribution Toolkit – final version,” Project MARVEL, 2023. <https://doi.org/10.5281/zenodo.8147109>. (Accessed 10th December 2023).
- [20] Innovation Radar <https://innovation-radar.ec.europa.eu>. (Accessed 10th December 2023).
- [21] Angular 2.0, Google. Available at <https://angular.io/>. (Accessed 10th December 2023).
- [22] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M. and Kalinin, A.A., 2020. “Albumentations: Fast and Flexible Image Augmentations”, Information, MDPI, vol. 11, issue 2, article 125, pp. 1-20. (DOI: 10.3390/info11020125).
- [23] Luca Zanella, et al., “Responsible AI at the edge: towards privacy-preserving smart cities”, Ital-IA 2022 Convegno del Laboratorio nazionale CINI-AIIS, Torino, Italy, 9-11 February 2022. (DOI: 10.5281/zenodo.5960431).
- [24] “D3.2: Efficient deployment of AI-optimised ML/DL models – initial version,” Project MARVEL, 2022. <https://zenodo.org/records/6821232>. (Accessed 10th December 2023).
- [25] “D3.6: Efficient deployment of AI-optimised ML/DL models – final version,” Project MARVEL, 2023. <https://zenodo.org/records/8147021>. (Accessed 10th December 2023).
- [26] Ultralytics YOLOv5 for object detection, instance segmentation and image classification, PyTorch. Available at https://pytorch.org/hub/ultralytics_yolov5/. (Accessed 10th December 2023).
- [27] “D4.6 - MARVEL's decision-making toolkit – final version,” Project MARVEL, 2023. <https://doi.org/10.5281/zenodo.8147077>. (Accessed 10th December 2023).
- [28] “D3.1: Multimodal and privacy-aware audio-visual intelligence – initial version,” Project MARVEL, 2022. <https://doi.org/10.5281/zenodo.6821318>. (Accessed 10th December 2023).
- [29] Raptis, T.P., Cicconetti, C., Falelakis, M., Kalogiannis, G., Kanellos, T. and Lobo, T.P., 2023. Engineering Resource-Efficient Data Management for Smart Cities with Apache Kafka. Future Internet, 15(2), p.43. (DOI: 10.3390/fi15020043).

- [30] Raptis, T.P., and Passarella, A., 2023. “A Survey on Networked Data Streaming With Apache Kafka”, IEEE Access, IEEE, vol. 11, pp. 85333-85350. (DOI: 10.1109/ACCESS.2023.3303810).
- [31] Kafka Message Broker, Apache. Available at <https://kafka.apache.org/>. (Accessed 10th December 2023).
- [32] “D5.2: Technical evaluation and progress against benchmarks – initial version,” MARVEL Project, 2022. <https://doi.org/10.5281/zenodo.6322699>. (Accessed 10th December 2023).
- [33] “D5.5: Technical evaluation and progress against benchmarks – final version,” MARVEL Project, 2023. <https://doi.org/10.5281/zenodo.10438311>. (Accessed 10th December 2023).
- [34] HORIZON 2020, “G. Technology readiness levels (TRL)”, Work Programme 2014-2015, General Annexes, 2014. Available at https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf. (Accessed 10th December 2023).
- [35] Common Platform Enumeration (CPE), NIST. Available at <https://nvd.nist.gov/products/cpe>. (Accessed 10th December 2023).
- [36] National Vulnerability Database (NVD), NIST. Available at <https://nvd.nist.gov/>. (Accessed 10th December 2023).
- [37] Common Vulnerability Scoring System (CVSS), FIRST. Available at <https://www.first.org/cvss/>. (Accessed 10th December 2023).
- [38] Common Weakness Enumeration (CWE), MITRE. Available at <https://cwe.mitre.org/>. (Accessed 10th December 2023).

11 Annex 1 – Data Model

This Annex presents the *Data Model that was developed to represent datasets*, along with their descriptions, metadata, and underneath snippets. The dataset representation is made in a JSON format and consists of two segments. The first segment contains general descriptive information. The second segment includes technical description based on the dataset's type (audio, video, and audio-video). This second segment contains also a table of the snippets that are part of the dataset. Each snippet entry has its own description. A JSON template for a *dataset description* is referred to the following code sample.

```
{
  "key": "Will be filled automatically",
  "datasetInfo": {
    "dataset": "<String Reference that points to the corresponding dataset name, e.g., MARVEL.GRN_Device1.Video.Y.Y.OriginalData.03-02-2022_1>",
    "category": "<String Reference that points to the type of data, e.g., Audio, Video, Img>",
    "dataProvider": "<String Reference that points to the data pilot, e.g. UNS, GRN, MT>",
    "description": "<A text-based description of the data>",
    "use_case": "<String Reference that points to the corresponding use-case, e.g. GRN1, GRN4, MT1, UNS2, etc.>",
    "keywords": "<A set of keywords separated with (,) describing the topic of the data>",
    "status": "<String Reference that denotes the dataset's status, e.g., In progress, Completed but private, or Completed and public>"
  },
  "audio-video": {
    "audio_bitrate": "<the audio bitrate>",
    "audio_sampling": "<sampling frequency in Hz>",
    "video_resolution": "<the video resolution>",
    "video_fps": "<video frames per seconds>",
    "annotation_video_software": "<the software type used for video annotation>",
    "annotation_audio_software": "<the software type used for audio annotation>",
    "annotation_video_ontology": "<the ontology used for video annotation>",
    "annotation_audio_ontology": "<the ontology used for audio annotation>",
    "device_id": "<points to file the gives device info>",
    "latitude": "<the latitude of the recording>",
    "longitude": "<the longitude of the recording>",
    "duration": "<audio-video duration in seconds>",
    "snippets": []
  }
}
```

The following code sample represents the JSON template for a *snippet entry* in the abovementioned table “*snippets*”.


```

{
  "key": "Snippet's key in the HBase",
  "publication_date": "<Date of data creation>",
  "duration": "<snippet duration in seconds>",
  "starts": "<snippet starting point in seconds>",
  "ends": "<snippet ending point in seconds>",
  "timestamp": "<the timestamp of data ingestion>",
  "datanodeHDFS": "Will be filled automatically upon ingestion to the Corpus",
  "urlHDFS": "The snippet's name, e.g., MARVEL.GRN_36.Audio.N.N.OriginalData.29-04-2022_9",
  "datasetID": "Will be filled automatically based on the 'dataset' field of the related dataset",
  "snippetFileName": "The file name of the main snippet file",
  "annotationFileName": "The file name of the annotations files",
  "inferenceResultsFileName": "The file name of the inference results file",
  "annotatorId": "List of annotators",
  "annotationSummary": "<type of events that refer to the corresponding snippet, parse the annotation file -> to place here>",
  "additionalEvents": "<extra annotation events>"
}

```

Also, a *naming format* for datasets and their snippets has been specified. For the *unique identification of each dataset*, the Corpus will follow the following specification:

DatasetID = <Project>.<DataProvider> [optional] _ <Device_id>.<Category (Video, Audio, or Video-Audio)>.<Annotated (Yes/No)>.<Anonymised (Yes/No)>.<Original data or Augmented><Date (dd-mm-yyyy)>.<Incrementing number starting from '1'>

For example:

- MARVEL.GRN_mike1.Video.Y.Y.OriginalData.03-02-2022_1
- MARVEL.MT_cam2.Video.Y.Y.Augmented.01-01-2020_1

The *unique identifier of each snippet* in the dataset will follow the following specification:

Snippet_D = DatasetID__<Incrementing number starting from '1'>

For Example:

- MARVEL.GRN_mike1.Video.Y.Y.OriginalData.03-02-2022_1

- MARVEL.GRN_mike1.Video.Y.Y.OriginalData.03-02-2022_1__1.mpeg
(actual snippet file)
- MARVEL.GRN_mike1.Video.Y.Y.OriginalData.03-02-2022_1__1.txt
(annotation file)
- MARVEL.GRN_mike1.Video.Y.Y.OriginalData.03-02-2022_1__1.score
(scoring file)
- MARVEL.MT_cam2.Video.Y.Y.Augmented.01-01-2020_1__20

DRAFT

12 Annex 2 –Letter of Intent (LOI) Template

[Full Name]

[Company]

[Street Address]

[City, St Zip]

[Optional – Email Address]

[Today's Date]

Letter of Intent (LOI) for Provision of Data availability and Free Technical Support for the MARVEL Data Corpus

Dear MARVEL Consortium,

I am writing on behalf of [Company], hereinafter referred to as "Partner," to express our keen interest in retaining the publicly available datasets and providing free technical support and consultancy services to MARVEL Data Corpus, hereinafter referred to as "Corpus", for a period of one (1) year after the completion of the MARVEL project – **01/01/2024 – 31/12/2024**.

The purpose of this Letter of Intent (LOI) is to outline the basic terms and conditions under which Partner is willing to offer its support and expertise to Corpus. This collaboration aims to enhance the overall performance, operation, security/privacy, and user experience of the Corpus.

Key Terms and Conditions:

1. Scope of Technical Support:

Partner agrees to provide technical support services, including but not limited to troubleshooting, bug fixes, system maintenance, and consultations related to Corpus, for a duration of **one (1) year after the end of the MARVEL project at 31/12/2023**.

2. Response Time:

Partner commits to responding to technical support or other requests from Corpus within **3 working days** of receiving the request during standard business hours, **9:00-17:00 CET**.

3. Communication Channels:

The parties will establish clear communication channels for submitting technical support or other requests and for regular progress updates.

4. Exclusions:

This agreement does not cover support for any customisations, integrations, or enhancements beyond the standard functionality of Corpus, unless explicitly agreed upon in writing.

5. Termination:

Either party may terminate this agreement with **30 days** written notice to the other party.

6. Confidentiality:

Both parties agree to maintain the confidentiality of any proprietary or sensitive information shared during the course of providing technical support.

This Letter of Intent is not legally binding and does not create any enforceable obligations on either party. It is intended as a preliminary expression of the parties' intention to negotiate and finalise a formal agreement.

Sincerely,

(Sign here for letters sent by mail or fax)

[Typed Full Name]

13 Annex 3 – User Satisfaction Studies

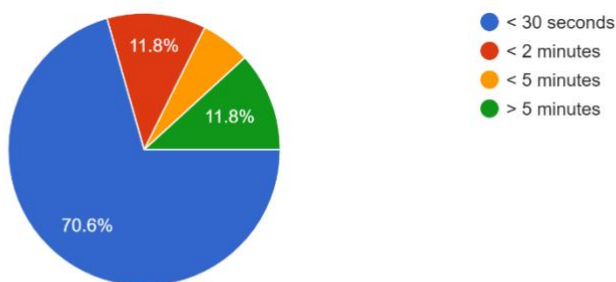
This Annex details the responses of the Corpus users, concerning the three evaluation studies that were conducted for ease-of-use and user-friendliness.

13.1 First Evaluation by Internal Users

13.1.1 Performance Evaluation section

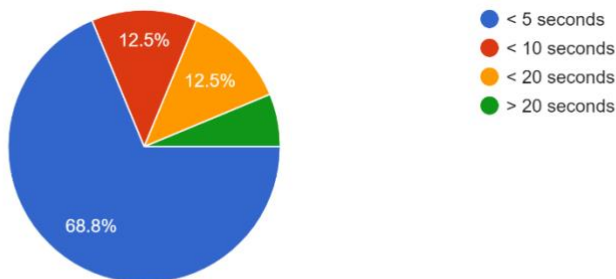
How much time required to download a demo snippet?

17 responses



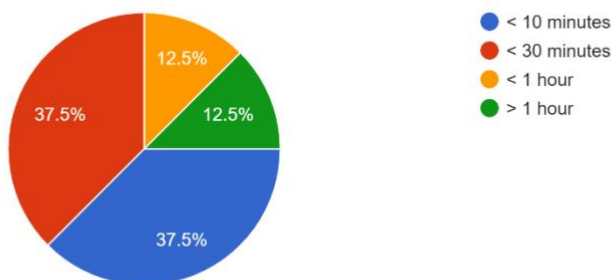
How much time required to search for a snippet?

16 responses



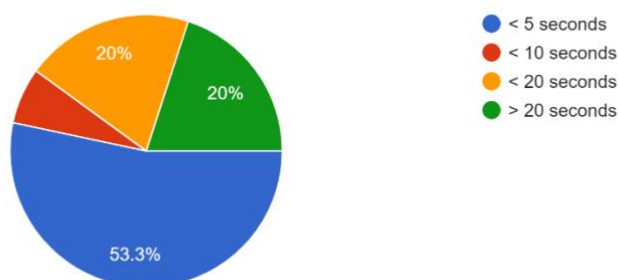
How much time required to download a demo dataset?

8 responses



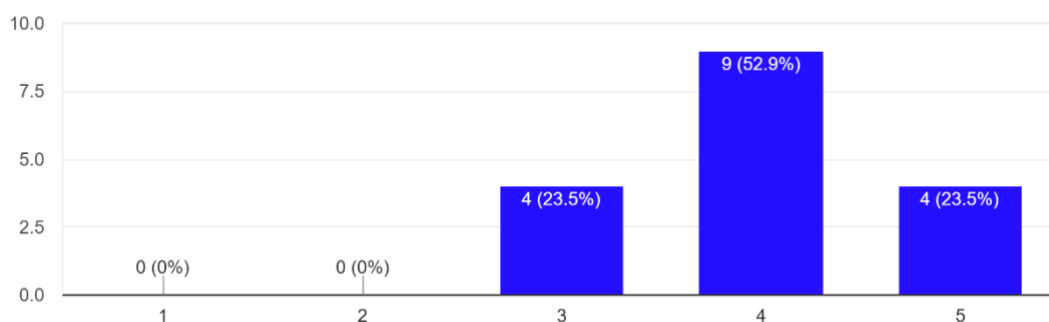
How much time required to search for a dataset?

15 responses



How would you rate the overall performance/responsiveness of the Corpus?

17 responses



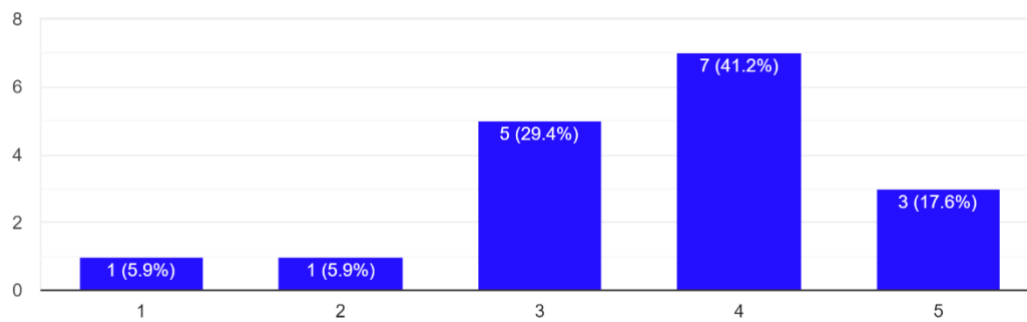
Other feedback/comments concerning the performance?

1. I did not find a way to download dataset from GUI.
2. We ca not download demo dataset
3. Data Corpus snippets slow to load and crashes.
4. Too much time to load the snippets dashboard. Check the name of the UNS use cases. I see 2. Use the full name as in the other use cases. Complex search does not seem to work. I am not able to download datasets, only snippets.
5. I am not sure if "Complex search" works. Also, I was able to download snippets but not a dataset.
6. UNS snippets seems to be only empty txt files. MT snippets could not be opened. From datacorpus.marvel-project.eu/snippets when I want to open GRN snippets, I got only JSON files. Only GRN snippets from https://datacorpus.marvel-platform.eu/snippets could be properly opened and viewed.
7. Search does not work for keywords

13.1.2 Usability Evaluation section

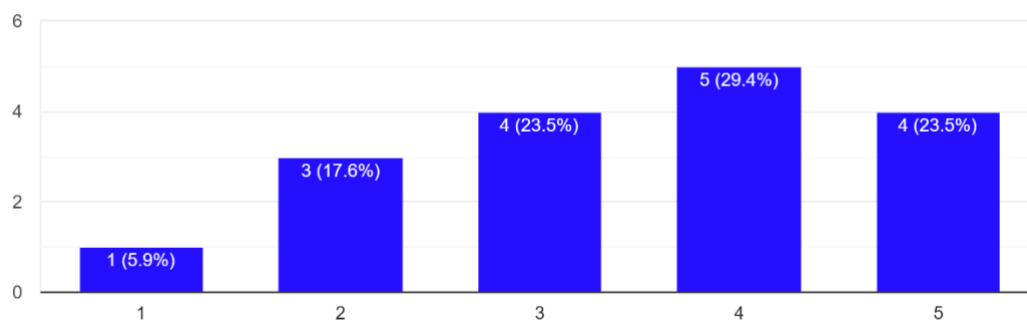
Was the information depicted for the overall datasets sufficient?

17 responses



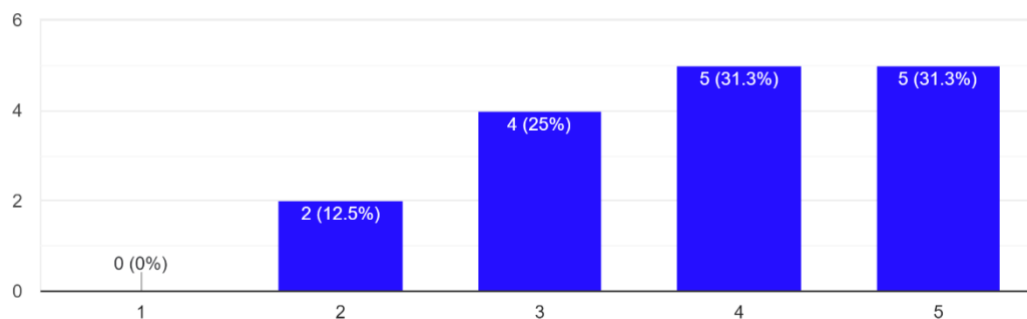
Was the search/filtering functionality for datasets sufficient?

17 responses



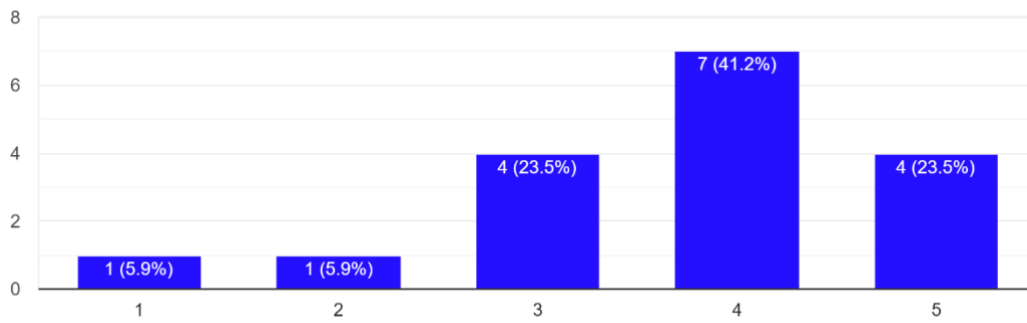
Was the detailed information depicted for a dataset entry sufficient?

16 responses



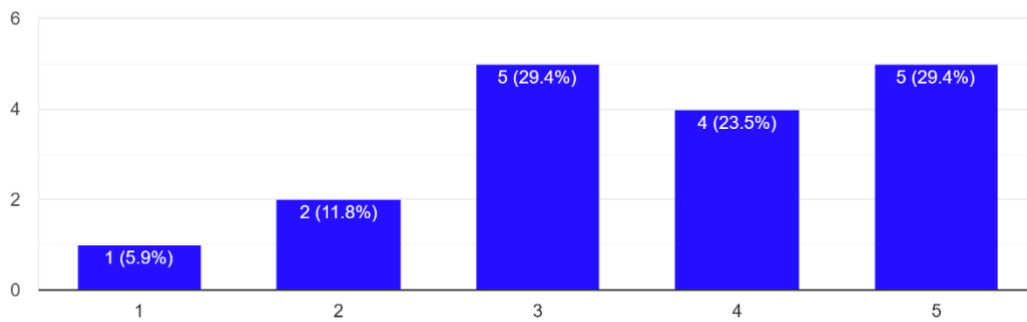
Was the information depicted for the overall snippets sufficient?

17 responses



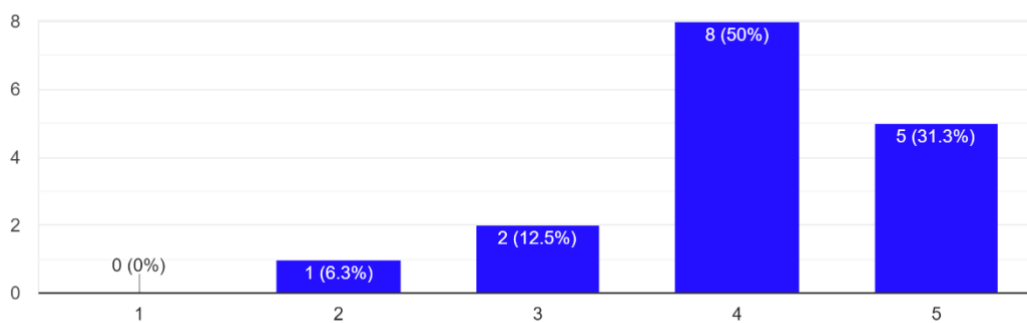
Was the search/filtering functionality for snippets sufficient?

17 responses



Was the detailed information depicted for a snippet sufficient?

16 responses



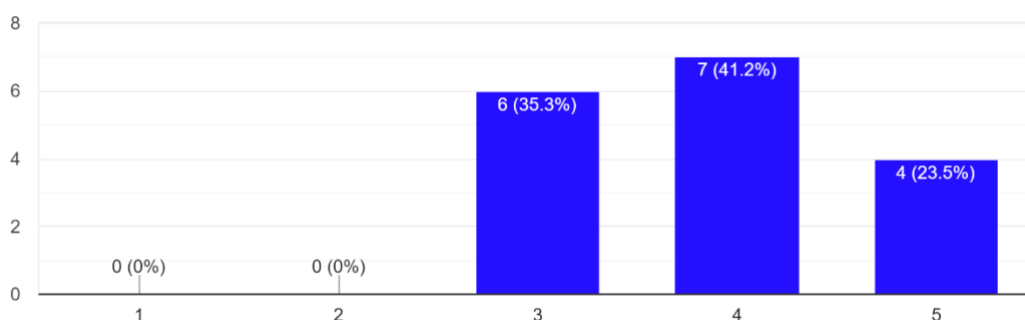
Other feedback/comments concerning the usability?

1. More information per overall dataset/snippet is needed to be shown in tables: amount of snippets per dataset, total length of dataset or snippet, is the data annotated. Per item details contains data only basic info, more detailed information is needed it to be informative for academic users.
2. (1) Complex search functionality was not working (after submitting search criteria, the results were not filtered. (2) it was not possible to download datasets. (3) it was not possible to view and/or download annotation files. (4) it was not possible to jump from to snippet files references within snippet files (cross-reference hyperlinks).
3. As mentioned before, too much time to load snippets dashboard.
4. Searching for a specific content or snippet is not simple if a user does not know what is in the Corpus, or who are the data providers, the data categories, etc.

13.1.3 User-friendliness Evaluation section

Rate the user-friendliness of the Graphical Interface

17 responses



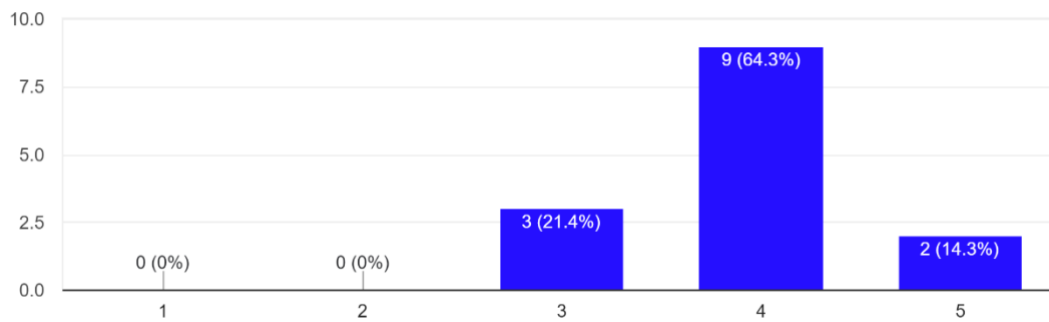
Other feedback/comments concerning the user-friendliness (e.g., bad design, missing features/functionality, etc.)?

1. There is no clear and direct connection to the MARVEL Website once you accept the policies, etc. Also, it might be a good idea to have the option for the dataset provider to create a profile with more info that will be available to the user when the user clicks on the provider's name.
2. (1) Drop down menus were not sufficiently indicated (e.g., arrow symbol). (2) cursor does not change to recognisable symbol when hovering over selectable items in the list of results.
3. Is it possible to download datasets or just snippets? Maybe you should clarify it.
4. The plots on the corpus composition are useful and other information about the composition can be made graphical, for easy and rapid access. Also, the filtering functionality can be made easier by providing pre-determined filters (e.g., the existing data providers, some informative category about the snippets in the corpus, etc.).
5. On graphs for snippets per provider or snippets per category, we could choose by one click a category and make a list of corresponding datasets (or snippets). However, search by entering a text is also fine for me.

13.1.4 Use and Exploitation Evaluation section

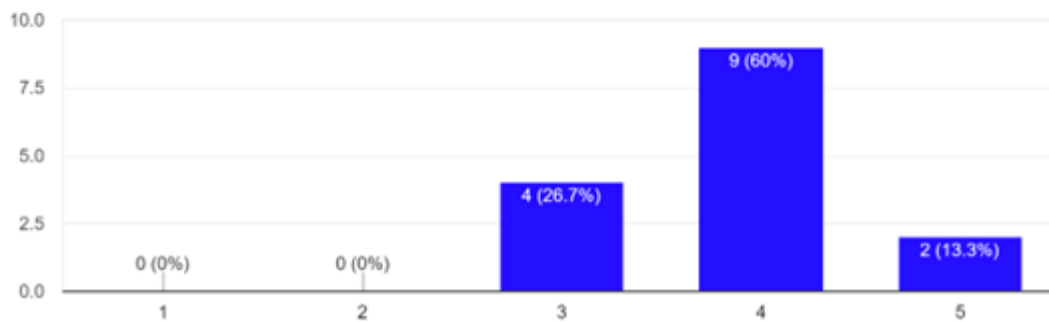
Was the data useful for you and/or your organization?

14 responses



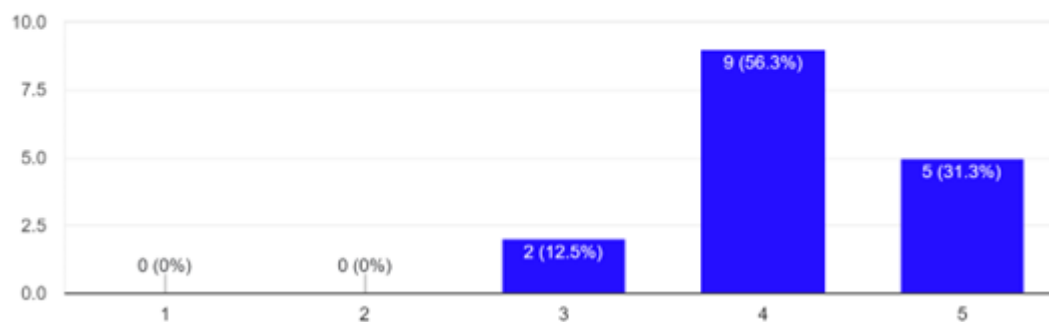
How likely is that you and/or your organization will use the MARVEL Data Corpus in the future?

15 responses



How likely is that you will recommend the MARVEL Data Corpus to colleagues and/or other entities?

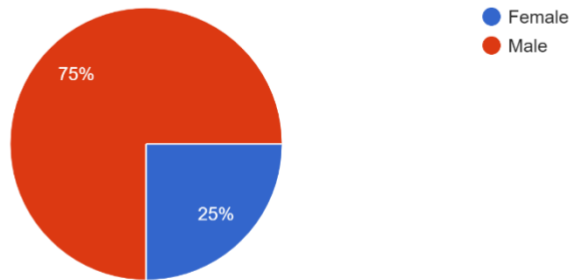
16 responses



13.1.5 Demographics section

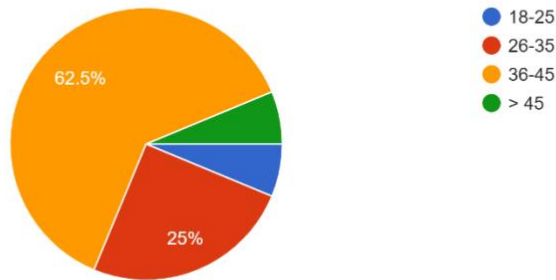
Gender

16 responses



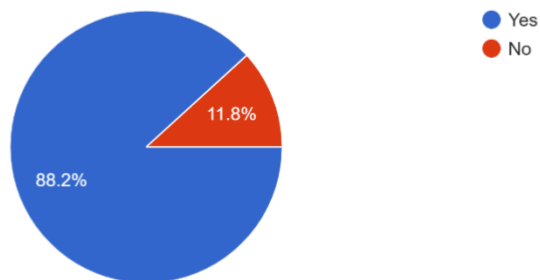
Age

16 responses



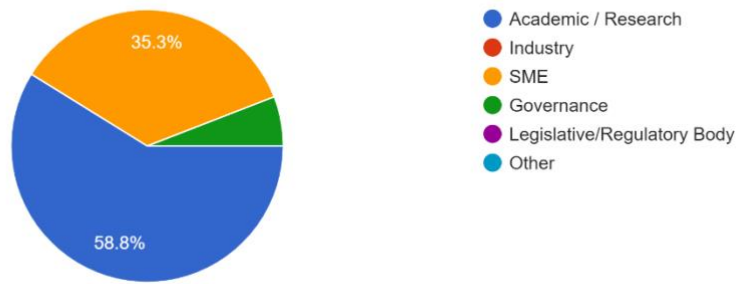
Are you member of the MARVEL Consortium?

17 responses



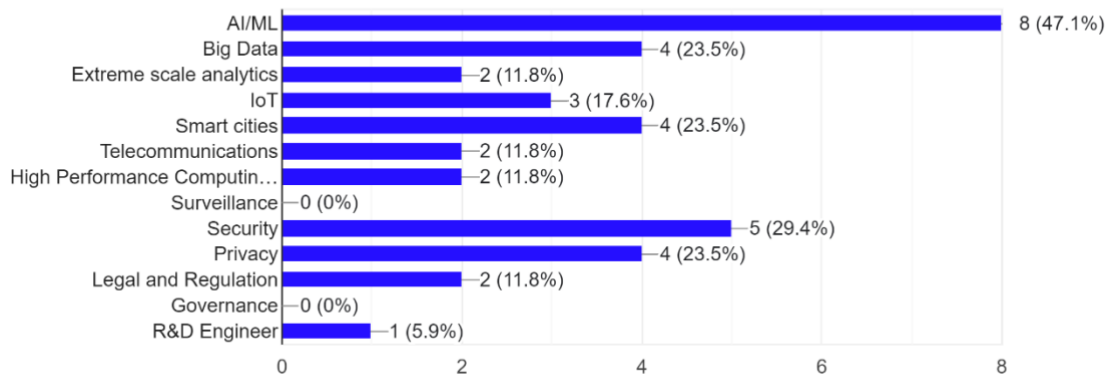
Organization

17 responses



Working field and expertise

17 responses

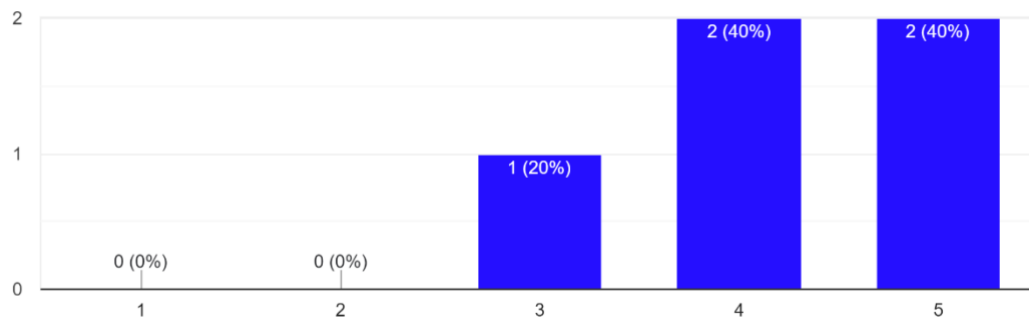


13.2 Second Evaluation by Internal Users

13.2.1 Performance, Usability, and User-friendliness Evaluation section

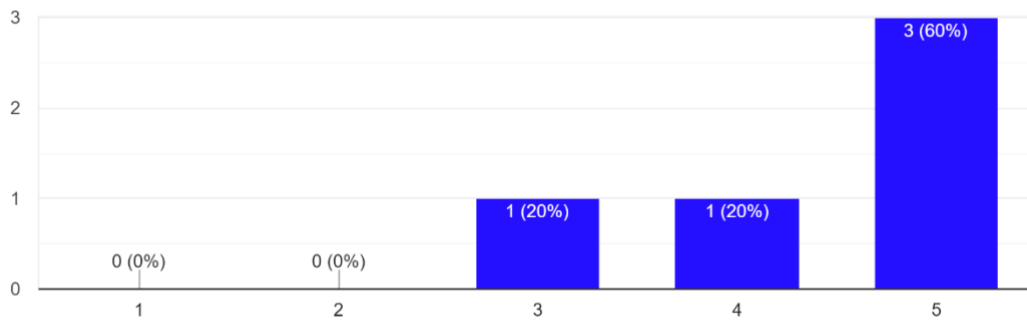
Rate the user-friendliness of the Graphical Interface

5 responses



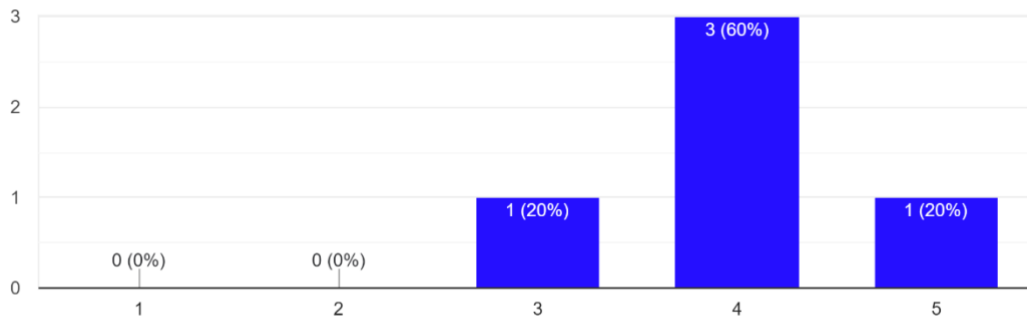
How would you rate the overall performance/responsiveness of the Corpus?

5 responses



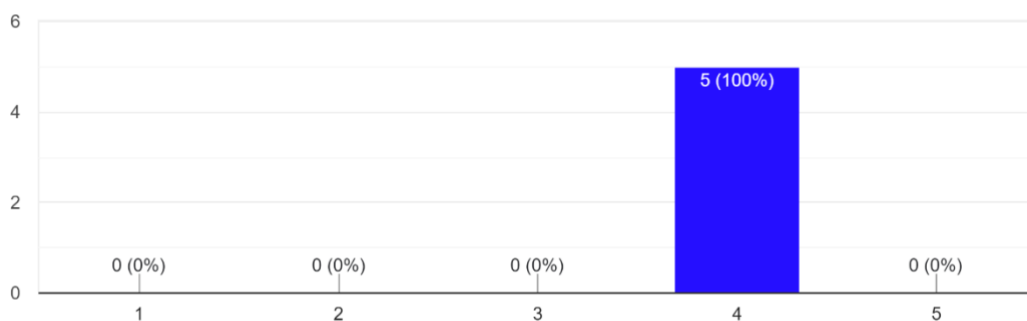
Was the search/filtering functionality for datasets sufficient?

5 responses



Was the information depicted for the overall datasets sufficient?

5 responses



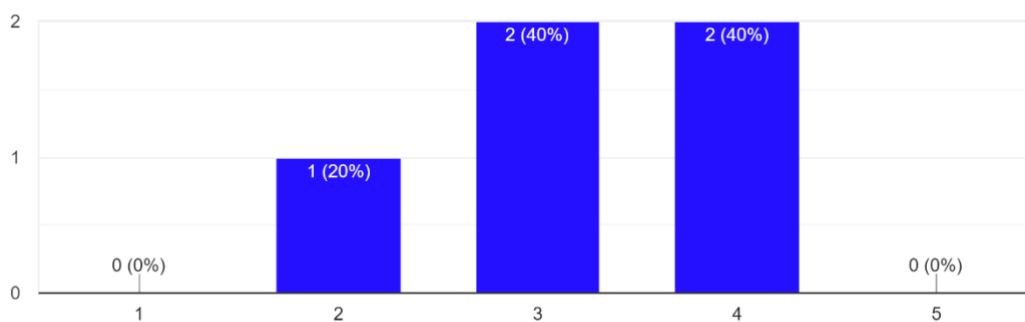
Other feedback/comments concerning the performance?

- Once the dataset is selected it takes time to load the dataset information, around 20 seconds. The downloading time depends on the size of the file, for bigger ones it took around 4 minutes.

13.2.2 Use and Exploitation section

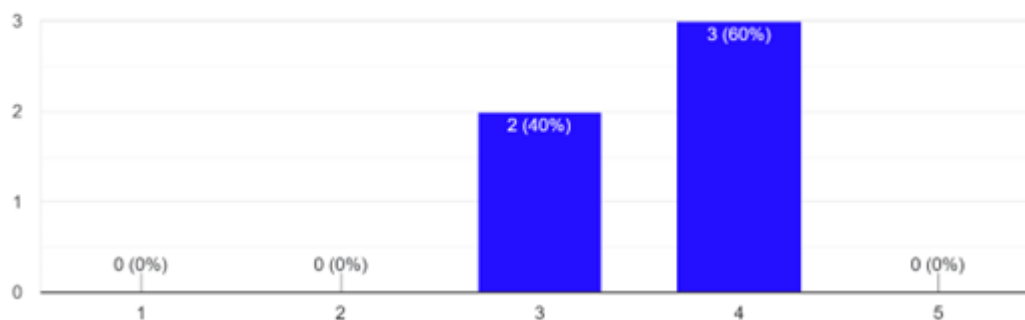
Was the data useful for you and/or your organization?

5 responses



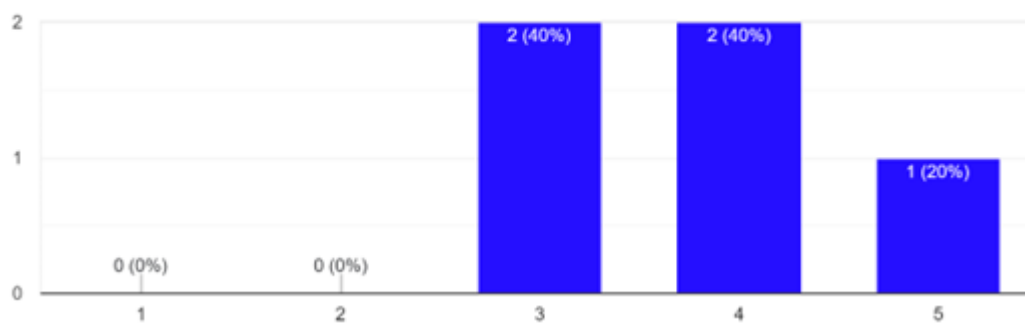
How likely is that you and/or your organization will use the MARVEL Data Corpus in the future?

5 responses



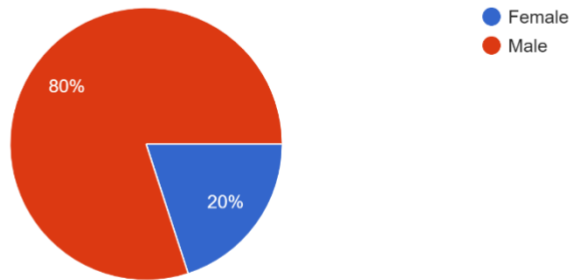
How likely is that you will recommend the MARVEL Data Corpus to colleagues and/or other entities?

5 responses

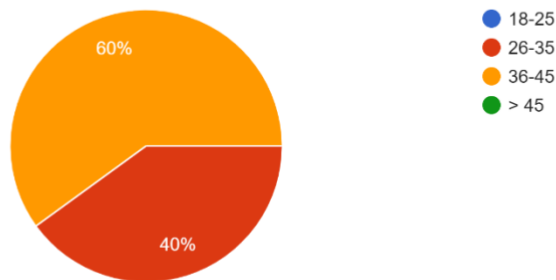


13.2.3 Demographics section

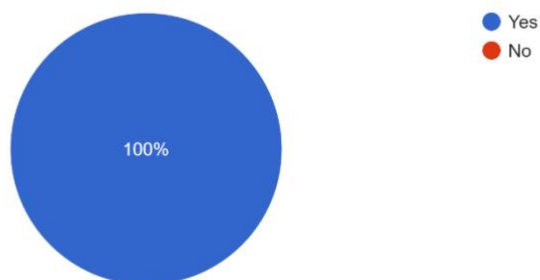
Gender
5 responses



Age
5 responses

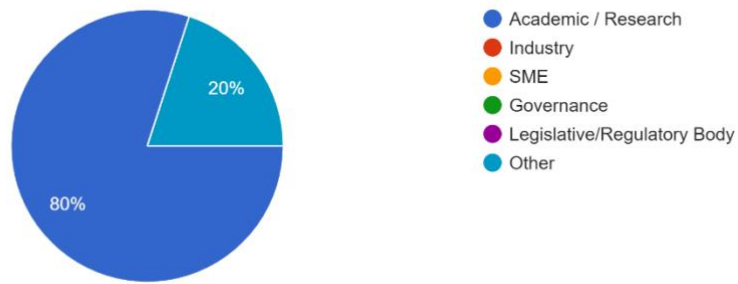


Are you member of the MARVEL Consortium?
5 responses



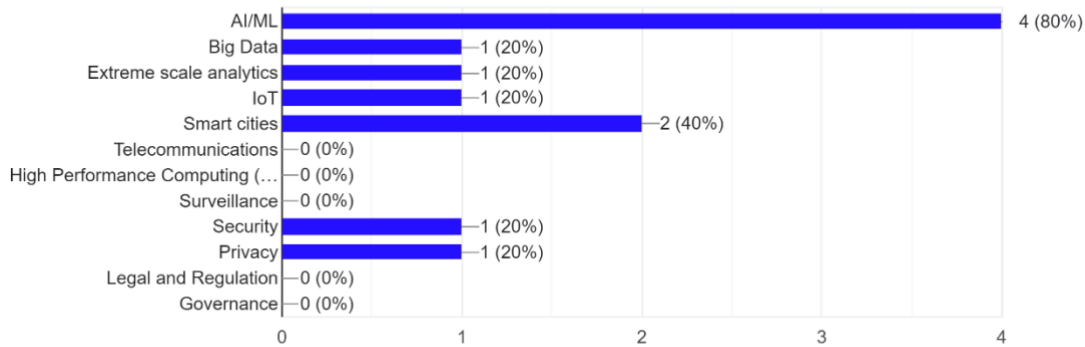
Organization

5 responses



Working field and expertise

5 responses

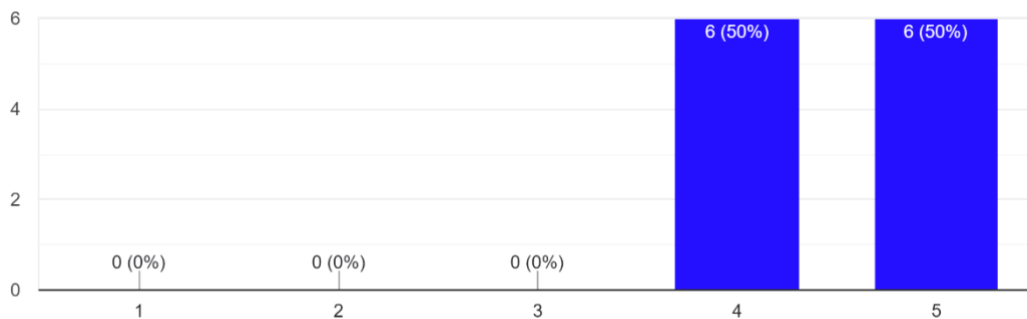


13.3 Third Evaluation by External Users

13.3.1 Trail and Evaluation section

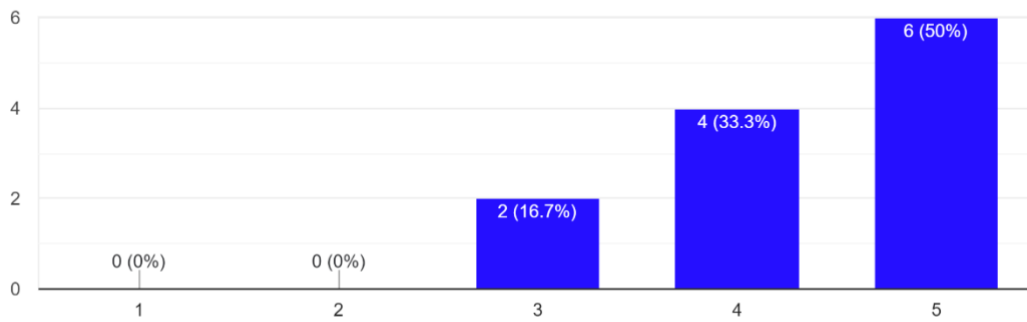
How easy was to use the Data Corpus?

12 responses



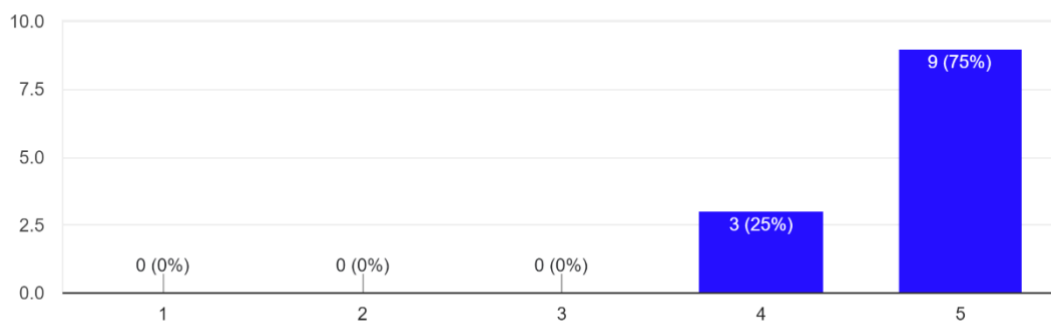
Was the information provided for the overall datasets and snippets sufficient to search and find the data that you need?

12 responses



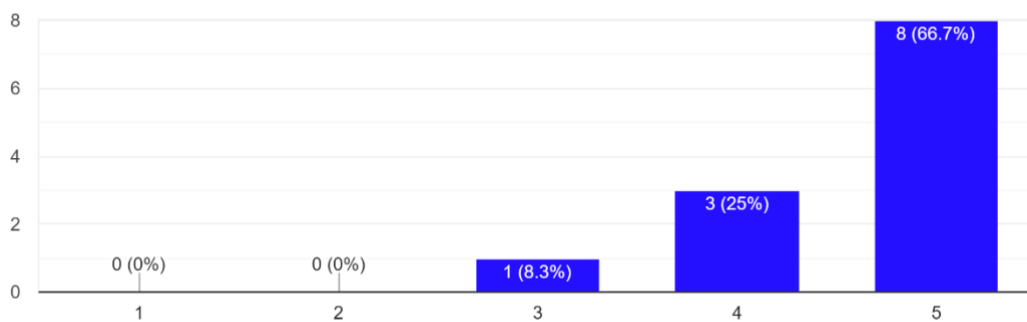
Was the overall functionality sufficient?

12 responses



How would you rate the overall performance of the Corpus?

12 responses

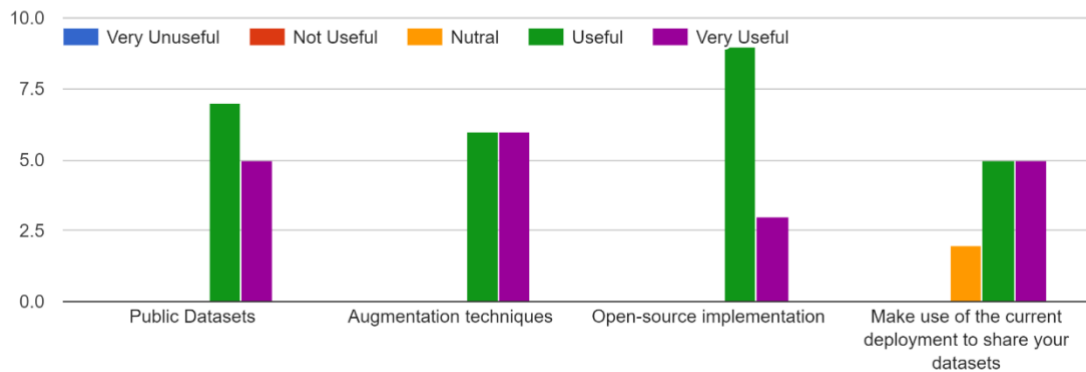


Any other feedback/comments?

- Very simple and easy.

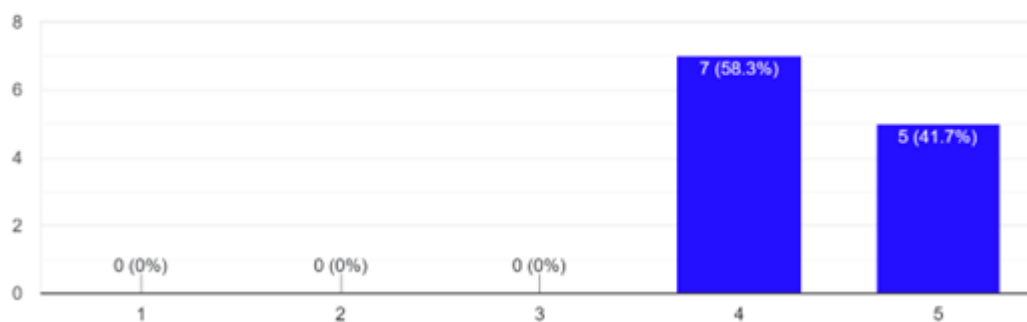
13.3.2 Use and Exploitation section

How useful do you find the Corpus features?



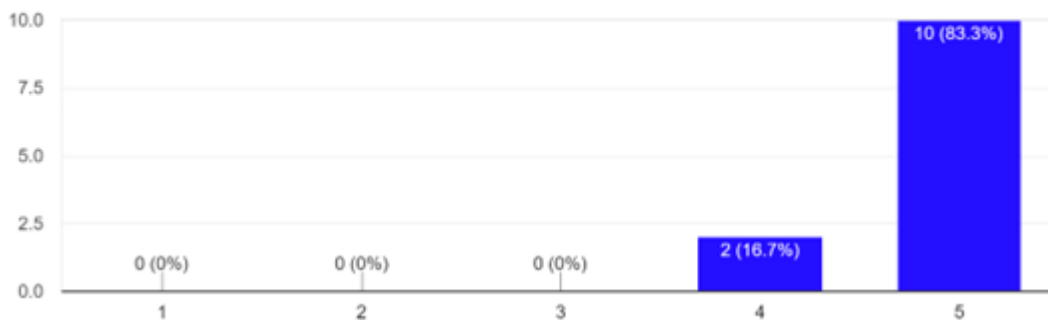
How likely is that you and/or your organization will use the public datasets in the future?

12 responses



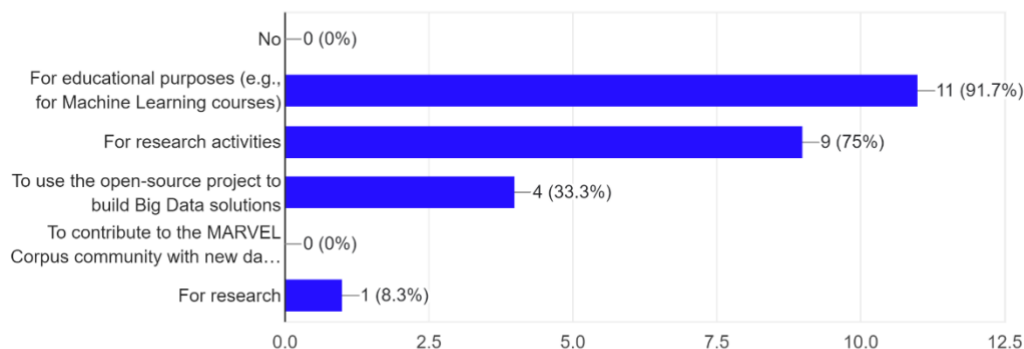
How likely is that you will recommend the MARVEL Data Corpus to colleagues and/or other entities?

12 responses



Would you be interested to use the MARVEL Data Corpus offerings?

12 responses



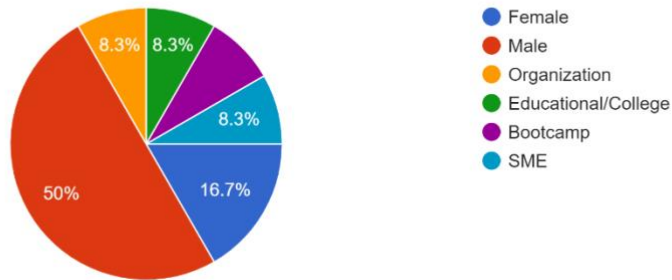
Please, specify if you are planning to utilise some of the Data Corpus offerings and how?

1. We are planning to use the public datasets within the under-graduate courses for Machine Learning and Pattern Recognition, as well as within students Thesis and academic research.
2. For research purposes on machine learning algorithms.
3. Creation of academic courses / workshops & hands-on hackathons
4. We could use the data in a future R&D project. We could also use the open-source implementation to build our own repository or upload our own datasets to MARVEL Corpus. Finally, another potential use of the Data Corpus is for training purposes.
5. These datasets could be used in ML and Security courses, dissertations, PhD research, etc.
6. Audio datasets can be used as future work in my PhD dissertation for the implementation of the spatial sound of external environments in a web 3D scene using ML algorithms.
7. Within courses and student research for activity recognition in public spaces, and monitoring of environmental aspects in smart cities
8. I will utilise some of the Data Corpus offerings in my research in Athens Metropolitan College.

13.3.3 Demographics section

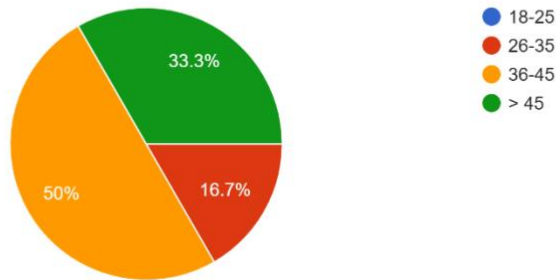
Gender

12 responses



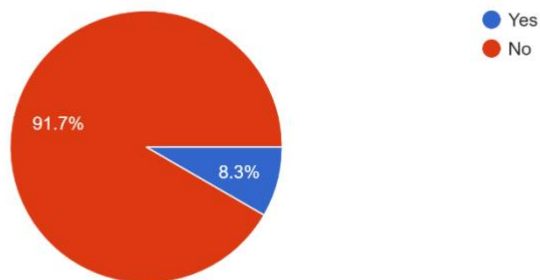
Age

12 responses



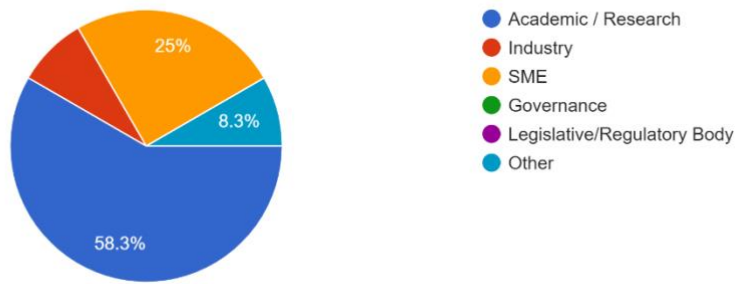
Are you member of the MARVEL Consortium?

12 responses



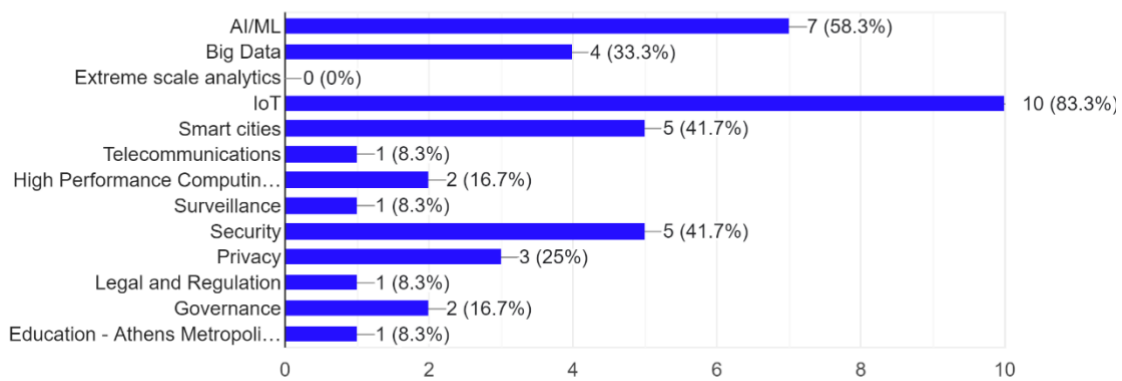
Organization

12 responses



Working field and expertise

12 responses



14 Annex 4 – Justification for the total Data Corpus size

This Annex provides the full justification for the lower performance concerning the “**KPI-O5-E1-1: More than 3.3PBs of data made available through a Corpus-as-a-Service**”, where the final Data Corpus reached a volume of around 1.1PB until M36.

When the proposal of MARVEL was submitted, we had set this KPI concerning the size of the Data Corpus that we would gather until the end of the project. The Corpus should store at least 3.3PBs of data. However, this goal was very overestimated, while on the other hand, several obstacles were underestimated.

The *main problems towards the estimation/calculation* of this threshold included that: i) already there were datasets from various partners as well as open datasets that could be shared and utilised for the project’s activities; ii) the collection of new data from pilots could start from day-1 of the project; iii) data of high resolution would be recorded; and iv) the infrastructure would be ready and the Corpus database and services implementation would be simple and available to run continuously without interruptions for most of the duration of the project (e.g., without considering stoppages during periods for implementation, deployment, testing, debugging, integration, etc.).

For the first part, *datasets that were available by some partners at the beginning of the project, were eventually not used due to raised ethical/legal issues and obstacles* (e.g., children carrying BLE beacons in their backpacks, raw datasets from recorded Maltese roads). Also, these datasets were not properly anonymised/annotated and the initial versions of the related MARVEL components that could curate these issues were implemented later in the project’s lifecycle, focusing the available effort on the processing of the new data that would be recorded by the pilots. Similarly, although we gathered several open datasets that someone could freely use for Machine Learning or other activities, after closely examining the legal aspects, we discovered that their current *open licenses were not permitting redistribution* in other repositories such as the MARVEL Corpus. Although individual researchers that were mentioned as contacts were positive to proceed, they did not fully own the datasets as they were produced under some projects or other activities of their organisations (which were not interested in sharing the data further). Alike with datasets that were produced by other collaborating EU projects like EVEREST and DAPHNE. Even if individual researchers were supporting these collaborations, after examining the official procedures that they had to follow, they were backing off. Thereafter, we examined the case to gather data from public cameras that are available on the Internet from several cities around the globe. Again, there were legal constraints as we could not receive permission from cities’ officials to store and publicly share such datasets.

Thereupon, we had to rely solely on the new data coming from the pilots. However, *MARVEL components had to be implemented and integrated before the gathering of data from the piloting environments became possible*. This was an issue that was underestimated when the proposal was written, and afterwards, when the project started, as the related Objectives and KPIs were considering that data collection was possible from day 1. It would need to collect, anonymise, annotate, and ingest around 3013GB (3TB) data per day for 1095 days in the row (3 years * 365 days). This was not the case as several components and technologies had to be built from scratch and work in a fine-grained manner, including elements for data collection, anonymisation and annotation, as well as the Corpus database service itself.

For the estimation of the value 3.3PBs, it was considered that we could utilise several stream sources that would capture continuous video of high and very high resolutions. Again, this was not the case that we actually faced. For the implementation of the rest of MARVEL's

requirements and goals, we had to accomplish real-time processing of the piloting streams. Involved operations included among others: anonymisation, AI-reasoning and activity inference, annotation, and augmentation. For all these actions, the higher the resolution, the more time is required. Therefore, one main strategy that is commonly adopted in smart city ecosystems is to use *streams of lower resolution with high compression and less frames per second*. This means that we had to *sacrifice the larger Corpus size to reach the real-time processing constraints*. *Nevertheless, the quality of the collected data was not downgraded*, as all valuable pieces of information can still be extracted by the gathered datasets. Thereupon, we managed to record and process several months of streams, but of lower size.

Another problem that we faced was the fact that the *privacy agency of the Municipality in Malta was reluctant to share the video streams* from the Maltese roads. Instead, they decided to *disclose only the relevant anonymised audio streams*, which were of a much lower in size than the video ones. After several long discussions and devoted effort by the involved partners, in the second half of the project, we managed to *come to an agreement with the city of Gozo in Malta* that could give us access to its public video streams and the permission to use the datasets for the Corpus objectives. *Additional deployment and integration efforts were required* by the technical partners to incorporate these streams as well. Then, in the second phase of the project, the MARVEL project started gaining attention in Italy, after some promotional events that were organised by the Municipality of Trento. The involved *Italian privacy authorities started re-evaluating the case, requesting details about the project and the processing activities* that had been performed until then. The privacy authorities in general questioned if the data coming from the cameras installed by the Local Police can be used for research purposes since there is a gap in the Italian Law framework. The *handling of such requests required additional effort from the technical partners* to collect and present all this information, as well as put stress on the team. A major side-effect of the achievement of the discussed KPI was the fact that *MT decided to stop the ingestion of new data until the situation had been resolved*. Also, it was requested to make all MT datasets confidential and establish a procedure to delete them in case that the privacy authorities were deciding so. This situation lasted for several months up to M36, losing the opportunity to gather original data and augment datasets for this period.

Typically, and based on the work plan, the Corpus tasks were meant to *start at M20* under WP2 and Tasks 2.3/2.4. Nevertheless, the involved partners delivered *core components and services much earlier* to support the various project activities. The main service APIs were successfully implemented for the MVP (M12) on-time, with a GUI being implemented for the 1st integrated version of the MARVEL platform (M18). The Corpus was used internally by MARVEL users to store piloting data. Several technical and security aspects have been resolved since then, and a first release of the Corpus for external users was ready for the second half of the project. Concerning the infrastructure that can store the required data volumes, PSNC had provided the servers to store up to 1PB of data. In the second half of the project, PSNC updated its infrastructure and provided servers that could expand their storage resources based on the project's needs and reach the 3.3PBs of volume. Of course, *refinements and updates in the implementation and infrastructure of the Corpus are still being performed as the Service started being used by internal and external stakeholders*, providing feedback for aspects like performance, user-friendliness, security/privacy, etc.

To sum-up the following **obstacles** were raised **along with the curative actions** that we took to handle them:

- Datasets that were initially available by several partners were not eventually used due to ethical/legal issues.

- After several iterations of discussions, the officials of these organisations did not give permission to use the data.
- Licensing of open datasets did not permit the redistribution by other repositories.
 - After searching related information from many datasets, we could not find datasets of large size that we could use. Even if individual researchers that were mentioned as contacts were willing to proceed, they did not fully own the datasets as they were produced under some projects or other activities of their organisations (which were not interested in sharing the data further).
- Similarly to the previous issue, licensing of datasets from other collaborating EU projects could not be granted.
 - Although individual researchers were at first positive about these collaborations, after examining the official procedures that had to be followed, they were backing off.
- Licensing of public camera streams did not permit storing data in public repositories without the official permission of the relevant municipalities.
 - No municipality was willing to sign such an agreement (except for Gozo in Malta which is mentioned below).
- The municipality of Malta decided to share only the audio streams and not the video ones that require more storage. Also, MT stopped recording and processing of its video streams during the second phase of the project.
 - After several discussions and opportunities that were examined, we managed to receive the permission from the city of Gozo to use their public streams. The produced Gozo datasets are within the largest datasets that we managed to form and significantly contributed to the final Corpus size.
- We had to process streams of lower resolution and frames per second to reach the constraints for real-time operation of other MARVEL components.
 - Nevertheless, the quality of the produced datasets was not downgraded as all valuable information can still be extracted.
- When the 3.3PBs volume was calculated it was taken into account that there would be available piloting data from day-1 and the ingestion to the Corpus would be continuous and uninterrupted, not considering properly the time required to implement, deploy, integrate, and test/debug the overall system.
 - Several components had to be built from scratch. We try to speed-up the development procedures and implement several core elements ahead of the original schedule.