Common European Data Spaces and Robust AI for Transparent Public Governance

CEDAR

Project acronym: CEDAR

Project full title: Common European Data Spaces and Robust AI for Transparent Public Governance

Call identifier: HORIZON-CL4-2023-DATA-01

Type of action: HORIZON-RIA

Start date: 01/01/2024

End date: 31/12/2026

Grant agreement no: 101135577

D2.1 Initial Data Catalogue and Data Preparation Methods

Document description: D2.1 provides the definition of the data sources used in the project, including synthetic data generation.

Work package: WP2

Author(s): Sophia Karagiorgou (UBITECH); Thodoris Semertzidis; Christos Chatzikonstantinou; Stefanos Demertzis (CERTH); José Miguel Blanco (UPM); Isabela Rosal Santos (KUL); Giulia Preti (CNT)

Editor(s): Irem Goymen (UBITECH)

Leading partner: UBITECH

Participating partner: CERTH, UPM, KUL, CNT

Version: 1.0 Status: Submitted

Deliverable type: Report Dissemination level: PU

Official submission date: 30/06/2024 Actual submission date: 28/06/2024

The CEDAR project has received funding from the European Union's Horizon Europe project call HORIZON-CL4-2023-DATA-01 funded project Grant Agreement no. 101135577

Disclaimer

The CEDAR consortium consists of the following partners:

No.	Partner Organisation Name	Partner Organisation Short Name	Country
1	Centre for Research and Technology Hellas	CERTH	Greece
2	Commissariat al Energie Atomique et aux Energies Alternatives	CEA	France
3	CENTAI Institute S.p.A.	CNT	Italy
4	Fundacion Centro de Technologias de Interaccion Visual y Comunicaciones VICOMTECH	VICOM	Spain
5	TREBE Language Technologies S.L.	TRE	Spain
6	Brandenburgisches Institut für Gesellschaft und Sicherheit GmbH	BIGS	Germany
7	Christian-Albrechts University Kiel	KIEL	Germany
8	INSIEL Informatica per il Sistema degli Enti Locali S.p.A.	INS	Italy
9	SNEP d.o.o	SNEP	Slovenia
10	YouControl LTD	YC	Ukraine
11	Artellence	ART	Ukraine
12	Institute for Corporative Security Studies, Ljubljana	ICS	Slovenia
13	Engineering – Ingegneria Informatica S.p.A.	ENG	Italy
14	Universidad Politécnica de Madrid	UPM	Spain
15	Ubitech LTD	UBI	Cyprus
16	Netcompany-Intrasoft S.A.	NCI	Luxembourg
17	Regione Autonoma Friuli Venezia Giulia	FVG	Italy
18	ANCEFVG – Associazione Nazonale Construttori Edili FVG	ANCE	Italy
19	Ministry of Interior of the Republic of Slovenia / Slovenian Police	MNZ	Slovenia
20	Ministry of Health / Office for Control, Quality, and Investments in Healthcare of the Republic of Slovenia	MZ	Slovenia
21	Ministry of Digital Transformation of the Republic of Slovenia	MDP	Slovenia
22	Celje General Hospital	SBC	Slovenia
23	State Agency for Reconstruction and Development of Infrastructure of Ukraine	ARU	Ukraine
24	Transparency International Deutschland e.V.	TI-D	Germany
25	Katholieke Universiteit Leuven	KUL	Belgium
26	Arthur's Legal B.V.	ALBV	Netherlands
27	DBC Diadikasia	DBC	Greece
28	The Lisbon Council for Economic Competitiveness and Social Renewal asbl	LC	Belgium
29	SK Security LLC	SKS	Ukraine
30	Fund for Research of War, Conflicts, Support of Society and Security Development – Safe Ukraine 2030	SU	Ukraine
31	ARPA Agenzia Regionale per la Protezione dell' Ambiente del Friuli Venezia Giulia	ARPA	Italy

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
0.1	03/05/2024	ToC released; and Partners allocation to sections	UBI, Sophia Karagiorgou
0.2	7/06/2024	First collaborative draft	All related partners
0.3	10/06/2024	Consolidated draft	UBI, Irem Goymen
0.4	12/06/2024	Input to section 2; end-to-end review and proof-reading	UBI, Sophia Karagiorgou
0.5	14/06/2024	1st draft – sent for internal review	UBI, Irem Goymen
0.6	21/06/2024	2nd draft regarding the internal review	VICOM, Amaia Gil Lerchundi VICOM, Francesco Zola ENG, Nicola Masi
0.7	27/06/2024	Final version	UBI, Irem Goymen
1.0	28/06/2024	Final manuscript submitted	CERTH, Thodoris Semertzidis

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.

Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

List of Tables

List of Figures

List of Terms and Abbreviations

ADA	Adaptive Discriminator Augmentation
AI	Artificial Intelligence
API	Application programming interface
AttnGAN	Attentional Generative Adversarial Network
AUC	Area under the ROC Curve
B2B	Business-to-Business
B2C	Business-to-Consumer
BT	Batch Table
CGAN	Conditional Generative Adversarial Networks
Charter	European Union Charter of Fundamental Rights
CNN	Convolutional Neural Network
CoE	Council of Europe
CVAE	Conditional Variational Autoencoder
DA	De-associative
DCAT-AP	Data Catalogue Vocabulary Application Profile
DDPM	Denoising Diffusion Probabilistic Model
DEs	Data Ecosystems
DF-GAN	Deepfusion Generative Adversarial Networks
DGA	Data Governance Act
DMP	Data Management Plan
DNN	Deep Neural Network
DP	Differential Privacy
DSA	Digital Services Act
DSSC	Data Spaces Support Center
dVAE	Discrete Variational Autoencoder
DX	Deliverable
ECHR	European Convention on Human Rights
ECtHR	European Court of Human Rights
EU	European Union
FID	Fréchet Inception Distance
G2B	Government-to-Business
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GLIDE	Guided Language to Image Diffusion for Generation and Editing
GPU	Graphics Processing Unit
GT	Generalized Table
I	Identifiers
IDS	International Data Space

IDSA	International Data Spaces Association
IEA	Interoperable Europe Act
IUN	Internet Uploading Number
LDM	Latent Diffusion Model
LDP	Latent Vector Differential Privacy
LED	Law Enforcement Directive
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
MX	Month
NLP	Natural Language Processing
NP	Non-perturbative
ODE	Ordinary Differential Equation
OOTS	Once Only Technical System
P	Perturbative
PixelCNN	Pixel Convolutional Neural Network
PixelVAE	Pixel Variational Autoencoder
PPDS	Public Procurement Data Space
PPTs	Privacy-Preserving Techniques
PRAM	Post Randomisation Method
PS	Pseudonymization
PSI	Public Sector Information
QI	Quasi-Identifiers
QIT	QI Table
RAM	Reference Architectural Model
RDF	Resource Description Framework
RNN	Recurrent Neural Network
ROIs	Regions of Interest
S	Sensitive
SAGAN	Self-Attention Generative Adversarial Networks
SotA	State of the Art
SSMI	Structural Similarity Index
ST	Sensitive Table
StackGAN	Stacked Generative Adversarial Networks
STFT	Short-Time Fourier transform
TED	Tenders Electronic Daily
TPU	Tensor Processing Units
TTS	Text-to-Speech
VAEs	Variational Autoencoder
vid2vid	Video-to-Video

VLAE	Variational Loss Autoencoder
VLOPs	Very Large Online Platforms
VLOSEs	Very Large Search Engines
VQ-VAEs	Vector Quantized Variational Autoencoders
WGAN	Wasserstein Generative Adversarial Networks
WPX	Work Package
xAI	Explainable Artificial Intelligence

Executive Summary

The CEDAR project aims to provide high-quality, analytics-ready, and open data for transparent public governance. Deliverable 2.1 Initial Data Catalogue and Data Preparation Methods, covers the first six (6) months of activities, focusing on identifying, collecting, and preparing data sources for the CEDAR project. It includes pilot partners' internal databases, public datasets accessed via APIs, web-scraped data and alignment with existing data catalogues and spaces. Where data is insufficient, or limited due to privacy concerns, then synthetic data generation techniques are employed. The report outlines data sources, preparation methodologies, synthetic data methods from the literature, and compliance with ethical standards and laws, setting the foundation for future updates and supporting CEDAR's data-driven governance objectives.

1 Introduction

CEDAR addresses the need for high-quality, high-value, analytics-ready, and open data for transparent public governance. In the present deliverable titled D2.1 Initial Data Catalogue and Data Preparation Methods, we report the activities performed in the first six (6) months of the project to identify and collect the data sources both coming from the technical and Pilot partners covering different modalities to set up the CEDAR Data Catalogue. In parallel, we have defined the methods to: (i) synthetically generate complex data sources; (ii) align with external Data Catalogues and data spaces from other EU and International initiatives. The data that has been currently collected comprise the initial CEDAR Data Catalogue which will be enriched in the upcoming period and will be continuously monitored to evolve as the project progresses.

As for now, the Initial Data Catalogue of CEDAR consists of:

- Pilot partners datasets and databases: All the pilot partners have made available a data dictionary and a representative sample of all the data sources that are linked with the problems they are currently facing and will be addressed within CEDAR.
- Public datasets: Publicly available datasets regarding tenders, bids and procurement processes have been made available through dedicated REST APIs or through scheduled services which collect overnight new data coming from open repositories.
- Data collected through web scraping and social media: Data which has been identified and will be made available to the CEDAR project from websites, social media and other data spaces.

For the cases where data has privacy restrictions, is unavailable or insufficient, synthetic data generation techniques will be employed. In this deliverable, we present a thorough literature review on the relevant methods per data modality along with our plans about synthetic data generation in CEDAR regarding data sources that are needed for analysis but are limited. By combining real-world data sources either contributed or collected in the project with synthetic data generation techniques, CEDAR will overcome data limitations and build robust Artificial Intelligence (AI) models with data of higher quality for effective analysis without disclosing sensitive information.

1.1 Positioning of the Deliverable within CEDAR

The deliverable D2.1, outlines the data sources populated in the CEDAR project in the first six (6) months, encompassing both real-world and synthetic data. This deliverable reports the activities conducted during the first six months of the project, detailing the identification, collection, and preparation methods for various data sources. The report aims to establish a foundational understanding of the data diversity and characteristics within CEDAR, ensuring transparency and accessibility to the entire CEDAR Consortium. D2.1 focuses on the following aspects:

- Data Sources Definition: Detailed descriptions of the data sources onboarded onto the project, including internal databases and existing datasets from the pilot partners.
- Data Preparation Methods: Explanation of the processes and methodologies adopted for preparing data to ensure it is high-quality, high-value, and analytics-ready.
- Synthetic Data Generation: Overview of the State-of-the-Art (SotA) techniques for generating synthetic data, including tabular, categorical, time-series, and multimedia data, along with gap analysis and methods applicable to CEDAR.
- Ethics and Legal Compliance: Discussion of the ethical standards, relevant laws, and anonymization techniques applicable to governmental data to ensure data privacy, security and compliance.

1.2 Structure of the Deliverable

The structure of the rest of this document is, as follows:

Section 2 Initial Data Catalogue of CEDAR provides an inventory of the initial data sources, including pilot-specific data, data from open sources and their preparation for the project.

Section 3 State-of-the-Art in Synthetic Data Generation outlines a thorough literature review of the existing techniques and identifies gaps in synthetic data generation, with a focus on methods applicable to CEDAR requirements for data.

Section 4 Ethics Standards and Legal Compliance addresses ethical considerations and legal requirements for handling data.

Section 5 Conclusion summarizes the deliverable.

2 Initial Data Catalogue of CEDAR

The following sections present our activities to identify, understand and aggregate the data sources and external data from open portals within CEDAR. Our effort initiated by revisiting the Pilots and identifying the relevant data for their challenges and problems, and we then continued by investigating and onboarding Data Catalogues and schemas from other open data sources and data spaces.

2.1 Initial Data Identification and Questionnaire

The identification of the data was streamlined through the population of a questionnaire across the CEDAR Pilots. This activity was completed by the setup of Data Management Plan (DMP) Registry (as reported in D7.2) also identifying data from the CEDAR partners. The result of the questionnaire circulated within the CEDAR Consortium contributed to understanding what kind of data we have currently available, define the plans to bring additional data sources and what other data are needed to achieve CEDAR objectives.

During the first six (6) months of the CEDAR project, through the questionnaire we managed to identify:

- General information about the pilots, the overall problem per pilot and details of specific problems per governmental institution, data modality and linked data to specific challenges;
- The relevance of data to specific challenges (e.g., numerical data, text data, images, etc.);
- The format, expected size, frequency of data updates and location of the data (e.g., spreadsheets, text files, databases, expected volume, local or decentralized storage);
- What attributes are needed to be extracted by the documents, spreadsheets or databases per specific challenge linked with the relevant analysis task.

Figure 1 depicts the results of the data identification through the questionnaires in CEDAR. Currently the CEDAR Data Catalogue is composed of four (4) Pilot Clusters with relevant data per pilot and from Open Data sources (i.e., Greek-Cluster). Each folder contains data of different modalities and formats along with a Data Dictionary with the necessary metadata describing the data schemas, attributes, specificities per Pilot.

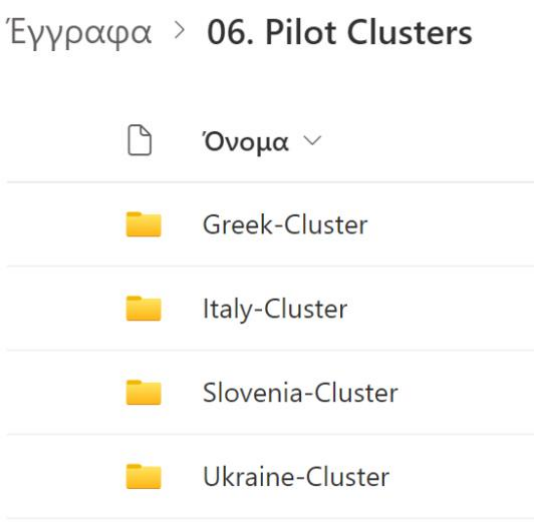

Figure 1. Data Collected for the CEDAR Pilots.

2.2 Data for the Italian Pilot

The data contributing to the CEDAR Data Catalogue from the Italian Pilot covers different problems and challenges. For this purpose, we have identified data coming from:

- Internal planning, evaluation and procurement processes (e.g., evaluation committees, evaluation criteria and evaluation reports);
- Tenders and rules about the rotation principle (i.e., linked with the criteria for eligibility and participation)
- List of invited and winning enterprises (e.g., evaluation / scoring criteria, blueprints, etc.);
- Ongoing contracts, public projects / works (e.g., services, construction, etc.) and execution variations;

- Suppliers profile (e.g., owners' information, average age of workers, etc.);
- Constructors profile, involved employees, on-site information and execution metadata (e.g., data for builders, involved workers, monthly working hours, beginning of building activity, equipment, external suppliers, etc.)
- Metadata about the tenders including the location, the type of work and the timing;
- Mass and social media; and
- Combinatorial data result sets among different platforms, e.g., Tenders Regional platform and Chamber of Commerce platform.

2.3 Data for the Slovenian Pilot

The data contributing to the CEDAR Data Catalogue from the Slovenian Pilot covers different problems and challenges. For this purpose, we have identified data coming from:

- Internal purchase orders (e.g., order identifier, ordered items, requirements, initiator, etc.);
- Tenders (e.g., tender identifier, ordered items, documentation, requirements, suppliers, necessary certificates or specifications, responsible, metadata, etc.)
- Bids (e.g., offer identifier, bid preparation date, bidder identifier, metadata and digital codes for each offered item, total price / offer / amount, delivery / submission date, responsible, metadata, etc.);
- Bidders profile (e.g., official business registry, history of their bids, success record in previous / similar tenders, network of contacts, bidding history of high-risk legal entities, criminal records and / or legal databases, etc.);
- Usage of public money (e.g., individual transactions, data on public procurements, e-invoices, etc.);
- Reports and past corrupt / fraudulent activities;
- Historical market rates or benchmark prices from external services;
- Public procurement data from other sources; and
- Public channels including social media and news.

2.4 Data for the Ukrainian Pilot

The data contributing to the CEDAR Data Catalogue from the Ukrainian Pilot covers different problems and challenges. For this purpose, we have identified data coming from:

- Tenders (e.g., identifier, title, amount, status, lots, etc.);
- Contracts (e.g., identifier, period, payment, date signed, contact points, metadata, etc.);
- Corruption registries verification, sanctions lists and tender participants connections check;
- Companies (e.g., representatives, connections among sibling / linked companies, etc.)
- Social media reputation; and
- Courts historical data, decisions on violation of laws and search.

2.5 Data from Other Open Data Sources

The Greek Transparency Program initiative begun in 1st of October 2010. All governmental institutions since then upload their decisions on the Internet with special attention to issues of national security and sensitive personal data. Each document is digitally signed and assigned a unique Internet Uploading Number (IUN) certifying that the decision has been uploaded at the "Transparency Portal." Following the latest legislative initiative (Law 4210/2013) of the Ministry of Administrative Reform and e-Governance, administrative acts and decisions are not valid unless published online. The main objectives of the portal concern:

- Safeguarding transparency of governmental decisions;
- Eliminating corruption by exposing it more easily when it takes place;
- Observing legality and good administration;
- Reinforcing citizens' constitutional rights, such as the participation in the Information Society;
- Enhancing and modernizing existing publication systems of administrative acts and decisions; and
- Making of all administrative acts available in formats that are easy to access, navigate and comprehend, regardless of the citizen's knowledge level of the inner processes of the administration.

UBITECH took the initiative to collect overnight decisions made by the Ministries aligned with the CEDAR Pilots in Greek and English to derive analytics to enrich the CEDAR insights coming from public expenses (cf. Greek-Cluster of Figure 1). The technological implementation model is based on an agile strategy of "open content" and "open architecture" allowing for the dissemination and re-use of Public Sector Information (PSI), providing the necessary tools for open and thorough access via the OpenDataAPI [98].

2.6 Other Data Spaces

Data spaces consist of trusted frameworks that manage the entire data lifecycle, encompassing various data models, metadata descriptors, ontologies for semantic interpretation, and data services for accessing, processing, and analysing data. Domain-

specific data spaces are currently being designed and deployed in vertical sectors, following specifications and reference frameworks that enable interoperability and compatibility. One such vertical sector with high impact on economy is governmental digital services covering public expenses, tenders and bids for transparent public governance.

Franklin et al. [85] originally introduced the concept of data spaces aiming to cover all data sources within an organization, irrespective of data model, data format or data location. Recently, the concept of data spaces has been revived due to the observed situation worldwide that clearly indicates that companies and organizations mostly operate as "data silos" or "data islands", without well-established procedures to facilitate data discovery, trusted and standardized data sharing, exchange and interoperability, leading to a waste of resources due to unnecessary and repetitive data-related operations [86], [87].

As such, the concept of data spaces has similarities with data marketplaces, which are online transactional locations or stores that facilitate the buying and selling of data, according to Snowflake [107]. Besides the provision of a trusted infrastructure for data discovery, sharing and exchange, data spaces comprise fully decentralized infrastructures that promote interaction in a federated or peer-to-peer way. In turn, this enables the participants or actors of data spaces to take the roles of data provider and/or data consumer, possibly after negotiation and settlement for a specific monetary value. Apart from data exchange, data spaces support the provision and consumption of services, thereby offering a fully-fledged platform for building next-generation applications.

As the public expenses domain related with the digitalization of governmental services is one of the fundamental pillars of the modern digital economy worldwide, the need for data spaces tailored to this domain is imperative. Having a dedicated data space for public spending allows for easier access to information by citizens, governmental bodies and observatory authorities. This transparency and accountability foster trust in public administration and spending and reduce the risk of public money misuse. Open data spaces for public expenses can encourage collaboration between government agencies, researchers, and businesses. This can lead to innovative solutions for managing public finances and delivering data-driven services and tools for promoting transparency, efficiency, and innovation in the digital economy.

To harness the value of data for the benefit of the European economy and society, the European strategy for data [99] of February 2020 set out the path to the creation of Common European Data Spaces in a number of strategic fields: health, agriculture, manufacturing, energy, mobility, financial, public administration, skills, the European Open Science Cloud. The green deal data space also stresses meeting the Green Deal's objectives as a key priority. We emphasize that data spaces for public administration are currently under development for the Legal and the Once Only Technical System (OOTS), while the Public Procurement Data Space (PPDS)[108] has recently launched in March 2024. We present in the following existing initiatives of Data Spaces that CEDAR plans to align with.

- The **International Data Spaces Association (IDSA) [109]**. It is a coalition comprised of more than 150 international organizations [110], which emerged in 2016 and has worked on the concept of data spaces and the principles that their design should follow in order to obtain value from data through sharing. The companies that comprise the IDSA represent dozens of industry sectors and are based in more than 20 countries around the world. As IDSA's mission is to drive the global digital economy, the main outcome of this work conducted by this coalition is to promote a Reference Architectural Model (RAM), called International Data Space (IDS) - RAM 3 [111]. The aim of this model is to standardize data exchange in such a way that participants can obtain all possible value from their information without losing control over it. This IDS-RAM is characterized by an open architecture, which is reliable and federated for cross-sector data exchange, facilitating sovereignty and interoperability.
- **Gaia-X [112]**, first introduced by the German and French Ministries of Economics in October 2019, is an initiative that unites various stakeholders to facilitate data and service sharing. Gaia-X enables data and service sharing amongst participants through the use of Federation Services. It can also merge different ecosystems by agreeing on a common operational model, i.e., the Gaia-X Operational Model, which is based on the defined Gaia-X basic concepts or the Gaia-X Conceptual Model. This concept is facilitated by the use of different planes, specifically the usage plane for technical interoperability, the management plane for governance, and the trust plane, which is supported by the common Gaia-X Trust Framework [113]. This framework provides the necessary set of rules that define the minimum baseline to be part of the Gaia-X ecosystem.
- The **Data Spaces Business Alliance**[114]. In September 2021, a collaboration was established between the Big Data Value Association, FIWARE Foundation, Gaia-X, and the IDS Association, resulting in the formation of the Data Spaces Business Alliance. The aim of this alliance was to create a common reference technology framework by converging existing architectures and models. The framework developed by this alliance enables businesses to securely share and exchange data, thus promoting innovation, growth, and competitiveness. The alliance unites a diverse range of businesses, IT providers,

research institutions, and other interested stakeholders from different sectors, all dedicated to the development and utilization of data spaces.

- **Open DEI** [115] project is an EU-funded project, which aims to detect gaps, encourage synergies, support regional and national cooperation, and enhance communication among the Innovation Actions implementing the EU Digital Transformation strategy. The cross-Industry Digital Platforms federation of the OPEN DEI project provides useful insights to the most relevant work in the field of Reference Architecture for building data ecosystems (e.g., data spaces) to support the digital transformation journeys in the four sectors targeted by OPEN DEI (i.e., manufacturing, agriculture, energy, and healthcare). This initiative aims to investigate a conceptual overview of the offered data ecosystems (Des) to guide their development.
- **Data Spaces Support Center (DSSC)** [116]. The Data Spaces Support Center is an initiative by the European Commission under the Digital Europe Programme to support the development and implementation of data spaces across various sectors. It aims to facilitate the creation of common European data spaces, which is a part of the European strategy for data. This is accomplished mainly by providing essential components that support the creation of data spaces, utilizing a range of technologies. In addition to that, the DSSC offers guidance on matters such as conceptual modeling (e.g., taxonomy, glossary) and the landscape of standardization (e.g. a compilation of standards).
- **iShare** [117] is an initiative that was launched with the goal of streamlining and standardizing data sharing, initially within the logistics sector. This initiative was set in motion by the Neutral Logistics Information Platform, a Netherlands-based foundation. The iShare project introduces a consistent set of agreements or norms for identification, authentication, and authorization. This allows entities within the logistics sector to share data in a more straightforward and secure manner, even with parties they have not previously interacted with. At the heart of this initiative is the creation of a trust framework that eases the process of data sharing among participants in the data ecosystem. The ultimate aim of iShare is to apply this framework along with existing initiatives (such as IDS, Gaia-X, FIWARE) to establish a solid European trust network for B2B data sharing.
- **deployEMDS** [118]. deployEMDS is a follow-up project co-funded under the EU Digital Europe Programme, designed to establish the necessary framework for interlinking existing federated ecosystems. This will be achieved by integrating common technical infrastructure and governance mechanisms for the upcoming implementation of the common European Mobility Data Space. The initiative fosters a wide European ecosystem of data providers and users, promoting the adoption of shared building blocks. Sixteen real-life use cases from nine EU countries have contributed to developing innovative services and applications in this area. The ultimate goal of deployEMDS is to accelerate sustainable and smart mobility, thereby reducing transport emissions, by focusing on implementing its infrastructure across the nine project sites.
- **Open EU Data Portal** [119]. Existing open data portals like data.europa.eu demonstrate the feasibility of federated data and could be integrated with future public expense data spaces. The commitment to open data research and learning is of paramount interest. It is a well-informed EU data space, empowered by timely and effective access to trustworthy information and knowledge and benefiting from all the opportunities this brings to society and the economy. The portal is a metadata catalogue. To foster the comparability of data published across borders, it presents metadata references using the application profile for data portals in Europe (Data Catalogue Vocabulary (DCAT-AP) [100]), using Resource Description Framework (RDF) technology. It provides translations of metadata descriptions in all 24 official EU languages [101] using machine-translation technologies (eTranslation [102]).
- **Simpl** [120] is an open source, smart and secure middleware platform that supports data access and interoperability among European data spaces. Simpl plays a major role in the creation of the Common European Data Spaces. These are data ecosystems where users in the same ecosystem share data in a safe and secure manner. Simpl gives data providers full control over who accesses their data in such data spaces.
- The **Public Procurement Data Space (PPDS)** [108] will connect European databases, including TED data [103] on public procurement, and national procurement data sets available in national portals. The concrete implementation and expected benefits are further explained in the Commission communication published on 16 March. The development of a truly integrated space for public procurement data will require a collaborative effort at EU, national and at the level of all public buyers across the EU. The PPDS will consist of 4 layers. *Data sources layer:* a federated network of connected data sources such as the European databases and procurement data sets available in national portals; *Integration layer:* the eProcurement ontology as common data format to create a harmonised data set; *Analytics layer:* capacities for data discovery, querying and data analysis to generate new insights, to make data-driven decisions and to set KPIs on priority policy areas; and *Client layer:* a user-friendly interface for the different users. Figure 2 depicts the different layers of EU PPDS.

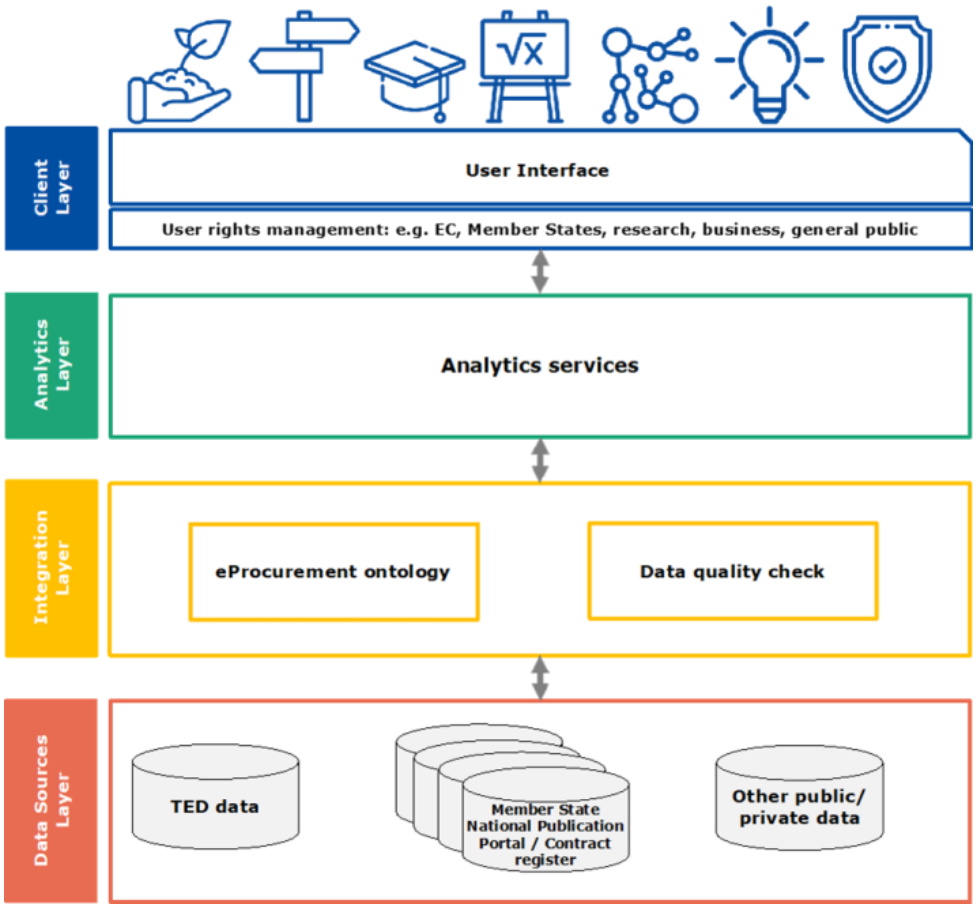

Figure 2. Layers of EU PPDS.

- **Other Common European Data Spaces** [121]. Data spaces in other important areas also exist such as media, energy, mobility, agriculture, finance, Green Deal, health, language, manufacturing, research and innovation, skills, tourism and cultural heritage. In D2.1, we concentrate on the most relevant data spaces to the CEDAR project, while we simply provide above an external reference for self-containment purposes. Together, the data spaces will gradually be interconnected to form the single market for data [104].

3 State of the Art in Synthetic Data Generation

This section presents a literature review regarding the algorithms and techniques for synthetic data generation targeting different data modalities.

Several works have been conducted on synthetic data generation due to lack of data, privacy issues, and to enhance the performance of the Machine Learning (ML) models. Data collection is specifically managed by the General Data Protection Regulation (GDPR) for personal data dictating policies and principles to privacy-preserving data access.

In addition, synthetic data generation is a growing approach to produce artificial data which resembles actual data source. Traditionally, most companies depend on the process of manual data gathering and labelling to meet the need of vast quantity training datasets, however this process can be insensible, costly and at times, not possible to achieve easily.

This is however achievable by synthetic data that make use of sophisticated mathematical models as well as other computational methods to create more real samples. As will be illustrated in the next sections, the synthesis of new data usually consists in reproducing the patterns, distribution, and dependency structures that are inherent to existing data sets. In the case of real data synthesis, synthetic data generators are capable of creating new samples which will be as probabilistically close to real samples as needed but at the same time containing controlled randomness and variation to still be distinct.

3.1 Tabular and Categorical Data

The creation of synthetic data is now an indispensable approach in several contexts like data augmentation, using synthetic data as a substitute for sharing actual data and as a means of preserving the private data's confidentiality. Therefore, the capacity to disregard the need for large amounts of quality data represents the relative importance and capability of synthetic data in data-driven applications, specifically in cases where real data is either inexistent, unavailable, or of low quality. Capturing the most recent literature on the subject matter of this overview, the key advancements and challenges in synthetic tabular and categorical data generation are examined.

Numerical values and categorical features on tables are types of big data that are frequently used in many applications. The method of generating synthetic tabular data has evolved greatly; especially in the years that deep generative models are in practice. In the following, we present an experimental study that focuses on using Generative Adversarial Networks (GANs) for creating synthetic tabular data. GANs consist of two neural networks: generator that will create fake data and discriminator that will differentiate the fake data from the authentic data. The generator is trained exactly to generate new data that they cannot distinguish from real data while the discriminator aids the generator in case of substandard data being produced. For example, several GANs specific for tabular data generation have been developed: Conditional GAN (CGAN), Wasserstein GAN (WGAN), which helps to increase the stableness and quality of generated data.

There are other types of generative models, one of which is called Variational Autoencoder (VAEs) that can also be used for synthetic tabular data generation. They get to know how to change the data into compact representation and vice versa, so that they are able to produce new data samples that look like the given data sample. However, VAEs have been used effectively in the generation of synthetic tabular data with both numerical and categorical predictors by incorporating appropriate encoders and decoders for each type of feature. Neural ordinary differential equations (ODEs) are specifically a flavour of deep generative models that can be applied to generate synthetic data in tabular format. They use ODE to mimic the behaviour of different data generation processes and they can produce replanted and realistic data. Many researchers have established that Neural ODEs are capable of capturing intricate dependency and correlation patterns in tabular data, which in turn translates to reliable generation of realistic synthetic data.

Secondary, categorical data including text, labels or any categorical features is another type of data used commonly in many applications. Data synthesis within categorical data has also received advancements in recent years of production. The main specific application of GANs is the generation of synthetic categorical data, which are different types of text and image data. It can produce many synthetic data samples which are close to the actual data and much more similar to actual data in realistic manner. Indeed, GANs have been extended to categorical data, and existing GANs that focuses on sequential data include SeqGAN and MaskGAN which is designed to generate sequences of meaningful categorical values.

VAEs have also been applied to synthetic generation of categorical data which is usually required in real world application. It enables them to learn how to create a compressed format and reconstruct the Categorical data from it and in this way, they can learn how to generate new data samples similar in nature to the original data. Some of the discoveries made in this work are; VAEs have been proven to work well in creating synthetic text data, the right encoder and decoder are used to match the

categorical nature of the data. It is a class of deep neural network that has been widely used in the synthesis of artificial discrete data, especially text. They can produce artificial data samples that are almost real and also would have a lot of variants from the actual data. It has been found that by using transformers, the distances and intricate repertoire of textual data can be captured at long range, thus creating high-quality synthetic data.

The creation of synthetic data is thus a widely applicable tool and has a range of usage scenarios in different domains. The above case meant that synthetic data can be used to enhance actual data, especially when information is limited or unbalanced. This increase in distinguishability might translate into improved performance of machine learning algorithms trained on the augmented data. Synthetic data can also be used to prevent individuals' identity in real data as through generating data which can have statistical characteristics but are not tied to any individual's identity. Synthetic data can help to disseminate personal information by releasing only the data in the healthcare system or financial system without endangering the privacy of those involved. Synthetic data can also be used for simulations or training exercises as real-world testing can be costly or may involve the risk of accidents, particularly in the case of autopiloted cars or drones, for example.

However, there are still several obstacles and prospects that have to be consider for the next steps in synthetic tabular and categorical data generation. The stages of data generation model evaluation also involve enhancing the quality of the created figures and developing reliable metrics for their assessment. Mean squared error (MSE) or mean absolute error (MAE), and other current metrics may not be enough to assess the characteristics of synthetic data. Such a sacrifice requires new measures that would reflect the statistical characteristics, usefulness, and anonymity of the obtained data. It is a critical issue as synthetic data generation models could potentially inherit the bias from the real data and thus might retain or even amplify the fairness and bias in machine learning models in the process of generating synthetic data. Thus, methods that need to suppress such biases have to be designed: for example, fairness constraints can be incorporated into data generation process, or debiased real data can be utilized for training.

Ten-scale models must successfully be extendable for synthesizing large-scale rates and distribution of data. Developing methods to scale up synthetic data generation models is essential for real-world applications, such as using distributed training or leveraging specialized hardware like GPUs or TPUs. Several cases arise where one needs to produce synthetic data with properties defined by some conditions or constraints such as label or other attributes. Engineering of accurate conditional synthetic data generation techniques stands as a promising avenue in the future. Evaluating the internal mechanisms of synthetic data generation models and why they produce specific data samples is crucial to trust Engineer and high-quality synthetic data set. It is the current research focus to build better machine learning interpretable synthetic data generation model.

In conclusion, synthetic tabular and categorical data generation has seen significant advancements in recent years, particularly with the rise of deep generative models. These models can generate highly realistic and representative synthetic data that is similar to the original data. However, there are still several challenges and future directions that need to be addressed, including evaluation metrics, fairness and bias, scalability, conditional generation, and interpretability. As the field continues to evolve, we can expect to see even more sophisticated methods for generating synthetic data that can be used in a wide range of applications, from data augmentation to privacy preservation to simulation and testing.

3.2 Time-series Data

The generation of synthetic time series data has been evolving in the recent past due to the impressive developments in deep generative models for temporal series data. They have made it possible to generate very realistic and genuine synthetic time series that can emulate the true data quite well both in terms of statistical property and dynamic behaviour.

Notably, the development of the GANs to create synthetic time series data is one of the key breakthroughs made in the field. GANs consist of two neural networks: a generator that produces synthetic data and a discriminator that distinguishes between real and synthetic data. The generator learns to generate time series that are indistinguishable from real data, while the discriminator helps the generator improve its quality. Several variants of GANs have been proposed for time series generation, such as WGAN and TimeGAN, which aim to improve the stability and quality of the generated data. The training of basic GAN is often constrained by instability; however, the WGAN addresses this problem by using the Wasserstein distance which provides more accurate convergence. In fact, TimeGAN uses recurrent neural networks to address the need for alternative models that capture the dependencies of the time series and the sequential nature of the data generation process, thereby yielding synthesized data that are more precise and reliable.

Another important development is the use of VAEs for synthetic time series generation. VAEs learn to compress and reconstruct time series data, allowing them to generate new data samples that are similar to the original data. VAEs have been shown to be

effective in generating synthetic time series with complex patterns and dependencies, particularly when combined with attention mechanisms and recurrent neural networks.

One of the key challenges in synthetic time series generation is ensuring that the generated data is not only realistic but also useful for downstream applications. To address this challenge, researchers have proposed methods for incorporating metadata or side information into the generation process. By conditioning the generation on relevant metadata, such as subject characteristics or experimental conditions, the utility of the generated data can be improved while still preserving the privacy of the original data [50].

Another important aspect of synthetic time series generation is the evaluation of the generated data. Researchers have proposed various metrics for assessing the quality of synthetic time series, such as resemblance to the original data, utility for downstream applications, and privacy preservation. These metrics help ensure that the generated data is of high quality and suitable for its intended use [51].

Although these aspects are clearer now, several issues and future stages can be identified in the context of synthetic time series generation. One problem is the expansion of the generative process to accommodate large sets and an abundance of data with nonlinear distributions. Furthermore, it is necessary to develop techniques protecting against providing unfairly generated data, which can be an extension of the existing biases present in the initial data provided to deep generative models if they are not corrected. The future work suggestion entails creating methods, which would allow creating conditional generation where the synthetic data generated would be in tandem with the requirement in the specific domain, incorporating domain knowledge in the generation of synthetic data to make them as realistic as possible, and adding more form of interpretability of the models used in generation of synthetic data in order to enhance trust in the process.

Summing up, the synthetic time series generation approaches can be considered as one of the most effective ways to cope with the problem of data insufficiency, personal data leakage, and data enhancements in different fields. Advanced methods like GANs and VAEs have allowed for generating samples indistinguishable from real life data series with high qualitative characteristics. Nevertheless, it is critical to continue investigating novel approaches and methodologies that would help address current and future limitations associated with the generation of synthetic time series data and improve their quality, usefulness and fairness.

3.3 Multimedia Data

3.3.1 Generic Multimedia data synthesis

Deep learning algorithms require a lot of high-quality data to train efficiently. However, real-world data is often of limited quantity or diversity, which can have a detrimental influence on model performance. To overcome this issue, two popular ways are synthetic data and data augmentation. From Figure 3, while both synthetic data and data augmentation aim to increase the size and diversity of the training data, they are not the same. Synthetic data is generated from scratch, while data augmentation uses the existing training data to create new examples

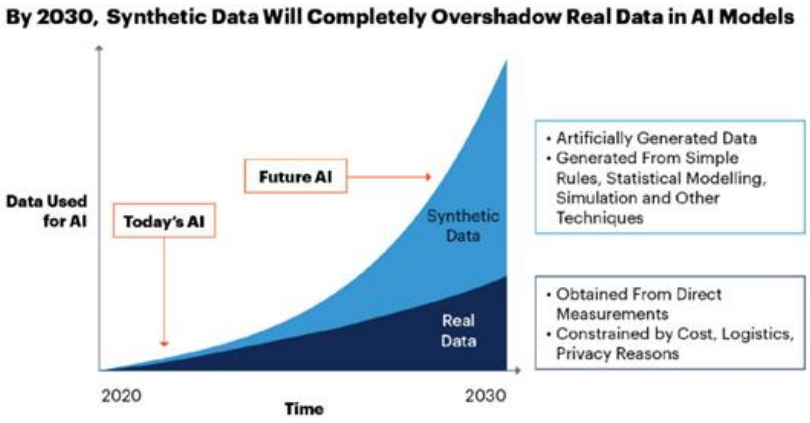

Figure 3. Synthetic Data Will Completely Overshadow Real Data in AI Models [105].

3.3.1.1 *Data synthesis*

Based on a survey by the Gartner company, by 2030, synthetic data will entirely outperform real data in AI algorithms. Therefore, the field of data synthesis is rapidly evolving, and new techniques are developed. In this section, frameworks, tools and literature algorithms about data synthesis will be presented.

3.3.1.1.1 Generative Adversarial Networks (GANs)

A widely used class of deep learning classes are generative adversarial networks (GANs). They were first introduced in [50]. In a GAN, two neural networks compete in a zero-sum game in which one agent's gain equals another's loss. The main concept of a GAN is built on "indirect" training via the discriminator, which may determine how "realistic" the input produced by the generator seems. Figure 4 represents the architecture of the data generators.

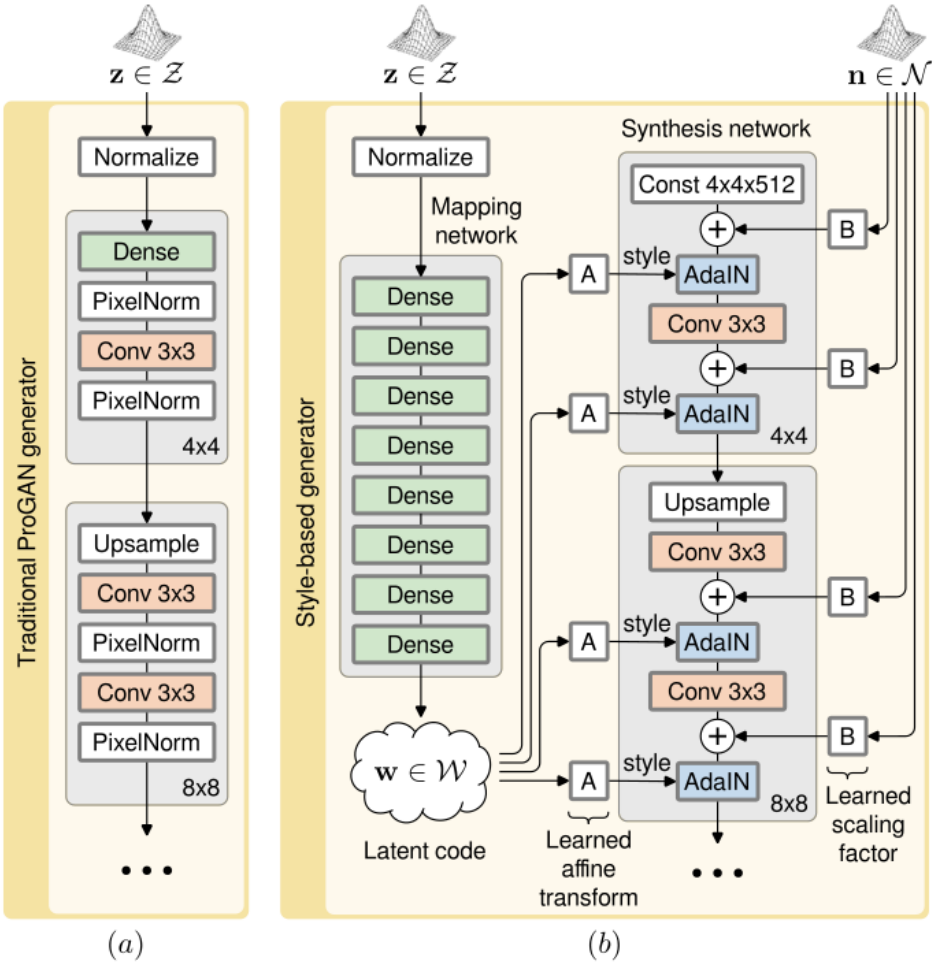

Figure 4. Generator Architecture for ProGAN (a) [52] and StyleGAN (b) [55].

StyleGANs [52],[53] are highly successful in image synthesis, producing realistic visuals. StyleGANs' well-behaved latent spaces that allow image manipulation directly on the feature space by vector operations [54]. It stemmed from the ProGAN [55] with a modified generator. There is a difference in the mapping network, mapping the latent vector into another vector in an intermediate latent space. The mapping network disentangles the latent space, resulting in higher linear separability and independent manipulation of the image features. A synthesis network also exists in StyleGAN. In this network, lower-resolution layers provide coarse elements like pose and facial shape, while higher-resolution layers generate finer details like hairstyle, colour scheme, and eyes.

StyleGan2 [56] architecture was developed to confront StyleGAN issues like blob-like artifacts and low shift invariance, causing features such as teeth and eyes to not line with the face when shifted. The problem was resolved by implementing a new generator with output skips and a discriminator with residual connections. Another problem that GAN training had to resolve, was the training of GANs with limited data. Therefore, the Adaptive Discriminator Augmentation (ADA) was developed allowing

data augmentation on GANs without the risk of augmentations being replicated by the generator [57]. Moreover, in order to improve StyleGAN robustness to translation and rotation, StyleGAN3 [58] was introduced with changes in the generator like redesigning the upsample and downsample filters, disabling the progressive growing, and filtering non-linearities. Despite their good performance, StyleGANs have a disadvantage: they need critical time and computer resources for training, making them difficult to duplicate. A more recent work is Style-GAN-XL that modified StyleGAN3-T architecture [58], reactivated the progressive training and employed class embedding information resulting in a larger network in depth and parameter count but with better results compared to previous methods on multimodal datasets.

One limitation of GANs has been the ability to generate images from diverse data distributions [59]. Zhang et al [60] assumed that that the convolutional layers commonly utilized in picture GANs contribute to the lack of diversity. The authors applied a self-attention mechanism in their network to capture long-range dependencies in pictures and improve results and named their network self-attention GAN (SAGAN) shown in Figure 5. In [61] the instabilities that occur when GANs are scaled up are studied. BigGAN was introduced, an updated stable architecture with more parameters and better results, demonstrating that GANs benefit from scaling.

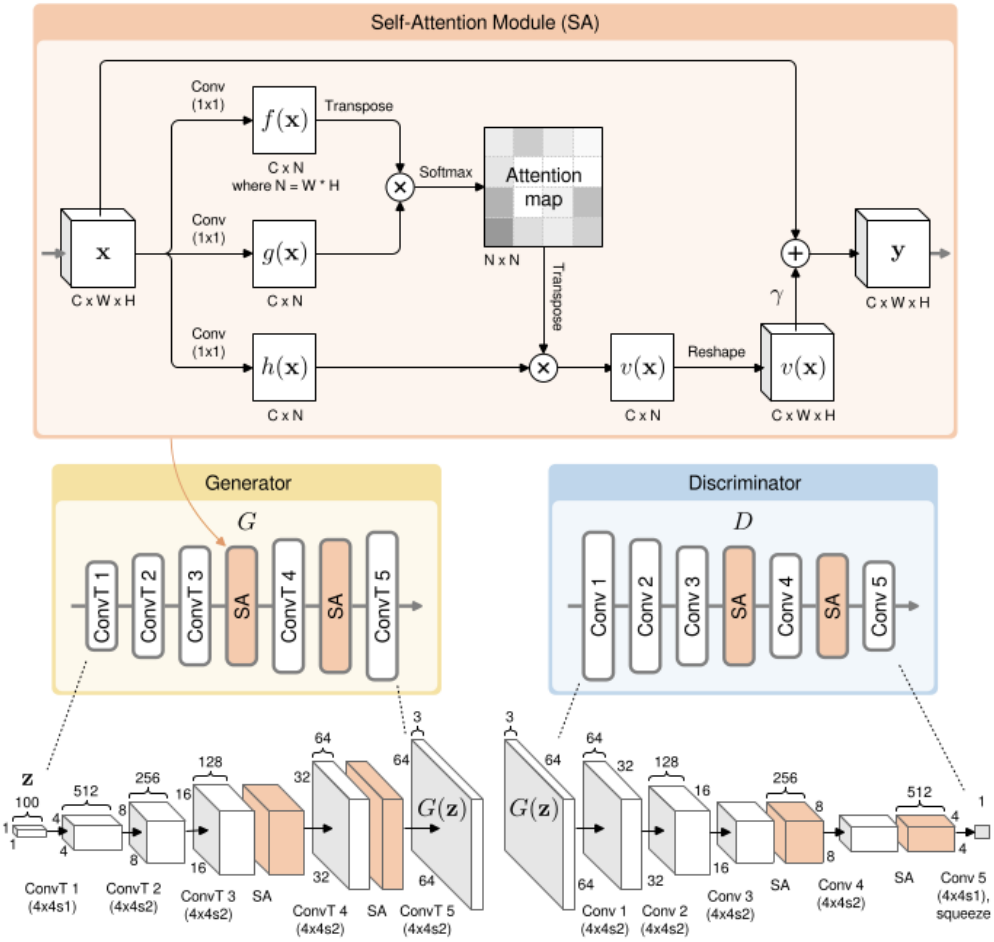

Figure 5. Self-Attention GAN (SAGAN) Architecture [54].

3.3.1.1.2 Variational autoencoders (VAEs)

The variational autoencoder is a data generation model based on variational Bayesian inference and it was introduced in [63]. Variational autoencoders map data to an ideal Gaussian distribution using an encoder. After sampling with a Gaussian distribution, the samples are sent into a decoder to generate reconstructed data.

Conventional VAE may create approximate input data but not particular types of data directionally. As a solution, conditional variational autoencoder (CVAE) was proposed [64]. In a CVAE, the model is given additional labels or data as conditional variables. These variables influence the encoding and decoding process, allowing the model to learn representations and generate data that are specific to the given conditions. CVAE's structure is like VAE's, making its computation and optimization methods comparable.

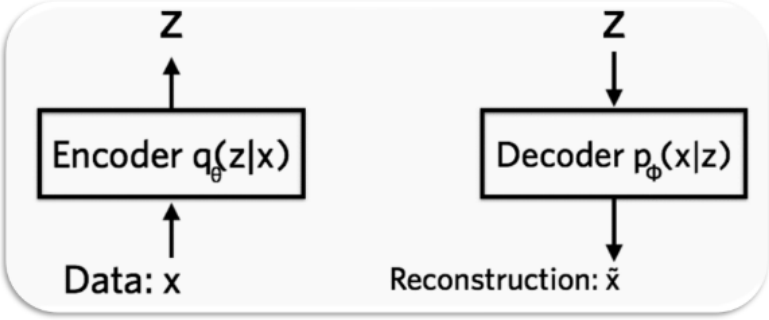

Figure 6. Abstract Representation of a Variational Autoencoder (VAE).

Researchers observed that combining VAE with an autoregressive network model improves its performance. The auto-regressive network model outperforms VAE due to its strong generating capability. Therefore, a combination of VAE and autoregressive network models (e.g. RNN) was proposed called Variational Loss Autoencoder (VLAE) [65]. Gulrajani et al. proposed a pixel variational autoencoder (PixelVAE) [66] that utilizes PixelCNN [67] to simulate an auto-regressive decoder for VAE. VAEs, which are conditionally independent between pixels, tend to create fuzzy samples, but PixelCNN, which models the joint distribution, produces clear samples. PixelVAE combines the advantages of both, providing meaningful potential performance and producing clear samples at the same time as shown in Figure 7.

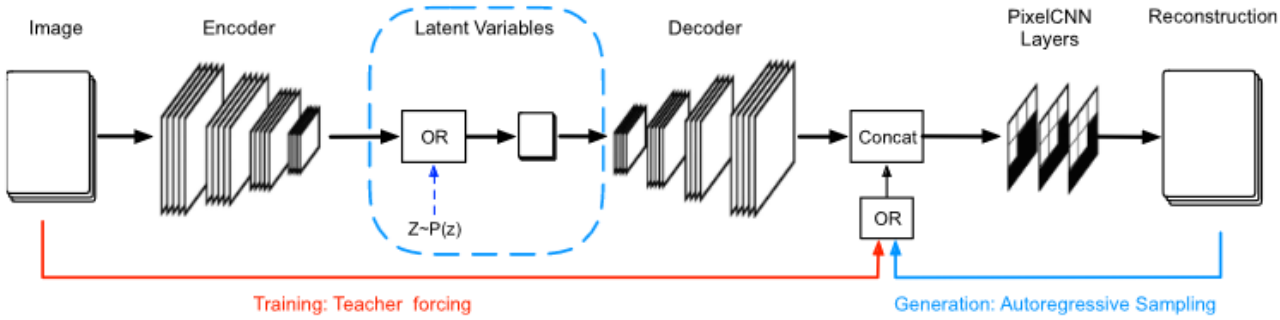

Figure 7. The Structure Diagram of Pixel Variational Autoencoder.

3.3.1.1.3 Text-to-image synthesis

Text-to-mage synthesis is the generation of images through text descriptions. More specifically, it is the use of computer tools to translate human-written textual descriptions (sentences or keywords) into visually similar representations. Word-to-image correlation analysis and supervised synthesis approaches were utilized to establish the optimal visual content alignment with text. It is a complex computer vision problem that has seen significant process in recent years. In Figure 8, presented architecture of text to image generation generally.

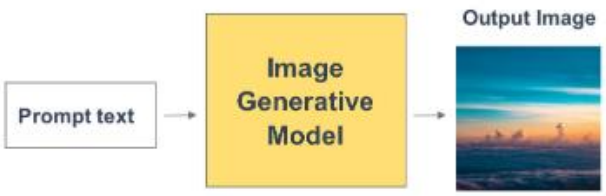

Figure 8. General Architecture of Text to Image Generation.

GAN models are widely employed for text-to-image synthesis. For instance, stacked generative adversarial networks (StackGAN) [68] developed a two-stage conditioning augmentation approach to increase the diversity of synthesized images and stabilize conditional-GAN training. For even more accurate text-to-image production, the attentional generative adversarial network (AttnGAN) [69] permits attention-driven, multi-stage refining. AttnGAN's attentional-generating network creates fine-grained picture features by focusing on key natural language terms. Not depending on any kind of entanglements among several generators, DeepFusion generative adversarial networks (DF-GAN) [70] can create high-resolution images using a single

generator and discriminator and allowing for a more thorough and effective fusion of text and picture information. In [71] transformer-based text encoders are combined with an advanced generator and create high-quality, text-aligned image generation. The model effectively generates pictures from complicated text descriptions, highlighting GANs' potential for text-to-image synthesis.

Diffusion models are also a class of models employed for text-to-image synthesis. Diffusion models are a class of generative models that learn to create data by starting with a random distribution and gradually refining it towards a target distribution over many steps. They are particularly known for generating high-quality images. Contrary to GAN-based approaches, those models use large-scale data to generate text-to-image conversions. However, the autoregressive nature of these techniques leads to high processing costs and error accumulation.

The VQ-Diffusion model [72] utilized vector quantized variational autoencoders (VQ-VAEs) and a conditional variant of the Denoising Diffusion Probabilistic Model (DDPM) to represent the latent space. In [73], the authors investigate CLIP guidance and classifier-free guidance as two separates guiding methodologies for the problem of text-conditional image synthesis. The proposed GLIDE model, which stands for Guided Language to Image Diffusion for Generation and Editing, was shown to be the most popular among people in terms of caption similarity and photorealism.

A breakthrough in the field of diffusion models were the DALL-E [75] models. The goal of the model is to train a transformer to autoregressively model the text and image tokens as a single stream of data. A discrete variational autoencoder (dVAE) compresses images to image tokens. The image tokens are concatenated with the text tokens and an autoregressive transformer models the joint distribution over those tokens. DALL-E 2 [76] aimed to create realistic images with higher resolutions by merging concepts, features and styles. The model has two parts: a prior that generates a CLIP image embedding from a caption and a decoder that generates an image based on the embedding. Another popular text-to-image tool is Stable Diffusion [101] that uses latent diffusion model (LDM). Stable Diffusion consists of the VAE, U-Net, and an optional text encoder. Compared to pixel-based diffusion models, LDMs dramatically reduced the requirement for processing while achieving a new state-of-the-art picture inpainting.

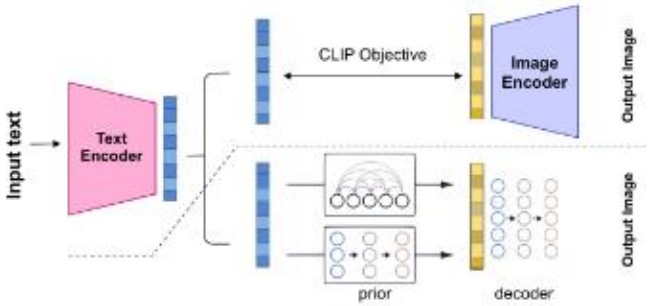

Figure 9. DALL-E 2 Architecture.

Figure 10. Images Generated by DALL-E 2 Given the Prompt: "a bowl of soup that is a portal to another dimension as digital art".

3.3.1.1.4 Video generation

The generation of videos from textual input poses a significant computational challenge. Nonetheless, recent advances in text-to-video artificial intelligence technology have demonstrated significant improvement in this field. Advances in realistic video creation and data-driven physics simulations are expected to accelerate the field's progress. In the next paragraph some of the SotA methods will be analysed.

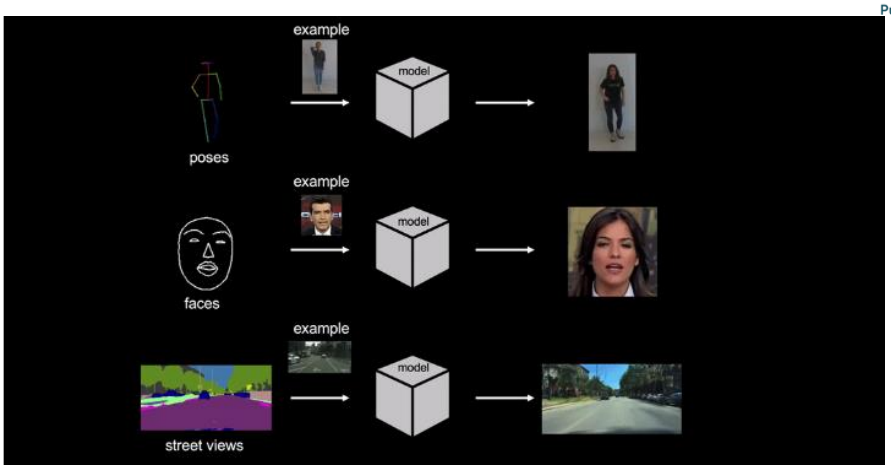

Figure 11. Video Synthesis Results of Few Shot vid2vid.

Few-shot video-to-video (vid2vid) synthesis [79] is an AI technique addressing the challenge of synthesizing photorealistic videos from a limited set of example images, contrary to traditional vid2vid models that require extensive datasets to train. This is accomplished via a network weight generation module that employs an attention mechanism, allowing the model to adapt to new input fast and effectively. Lumiere [77] is a text-to-video diffusion model that aims to create realistic and cohesive motion in synthesized videos, addressing a key challenge in video synthesis. The model uses a Space-Time U-Net architecture to generate the whole temporal duration of the video in a single pass. It employs spatial and temporal down- and up-sampling, as well as a pre-trained text-to-image diffusion model, to produce a full-frame-rate, low-resolution video over different space-time scales as presented in Figure 12. Stable Diffusion [78] has the benefit of producing high-quality images. This capability is based on improving throughput through several diffusion steps, increasing control by progressively adding noise and facilitating image synthesis. Furthermore, its capacity to generate realistic and subject-specific images can be better evaluated by assessing the model's subject-specific training results on custom datasets.

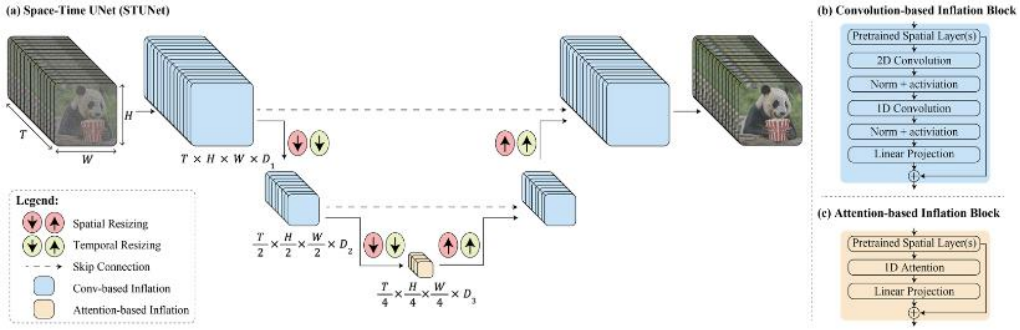

Figure 12. Space-Time UNet (STUNet) Architecture.

3.3.1.1.5 Online tools for video and image generation

Apart from the research that has been performed on the field, some online tools also exist for video and image generation. DALL-E, Midjourney, Sora and Stable Diffusion.

DALL-E has been created by OpenAI company. Its 3rd version is available online [80] and allows for creation of realistic images and art from a description in natural language (prompt). There is an API available, and it is also integrated in Microsoft products such as Bing, Microsoft Edge and Skype. An example of generated image with DALL-E can be found at Figure 14.

Another OpenAI product is Sora [81], a text-to-video model that can generate videos up to a minute long based on a user's prompt. It is based on text-conditional diffusion models trained jointly on videos and images of variable durations, resolutions and aspect rations. Videos are turned into patches, after they have been compressed into a lower-dimensional latent space. Spacetime patches are extracted, which function as transformer tokens. Sora is not yet publicly available for video generation, but this may change throughout the project. The architectural figure of Sora is represented in Figure 13.

OpenAI is not the only company that produces those kinds of models. Another online tool is Midjourney [107] created and hosted by the research lab Midjourney, Inc. Similar to the previous tools, Midjourney generates images from natural language descriptions (prompts). As things stand, it can be assessed through a discord server, and it requires a premium account.

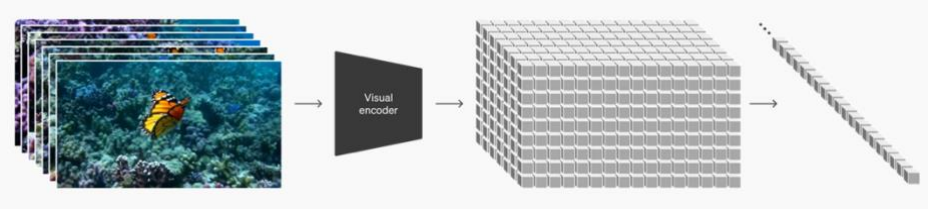

Figure 13. Sora Architecture: Videos Compressed Into a Lower-Dimensional Latent Space which is Decomposed into Spacetime Patches.

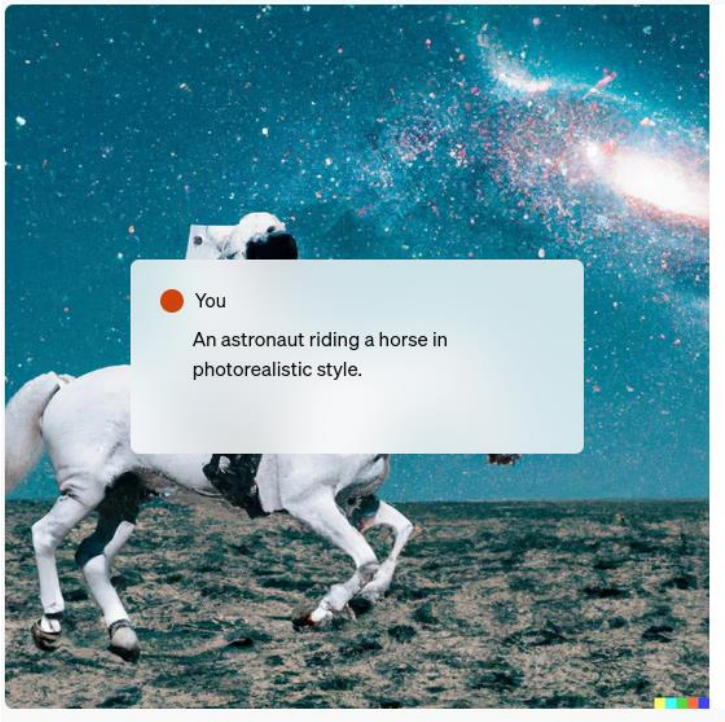

Figure 14. Image Generated from DALL-E 3.

Last but not least is the Stable Diffusion [82] online tool, a latent text-to-image diffusion model capable of generating photo-realistic images. It is available for free but with a limit on the number of prompts and images.

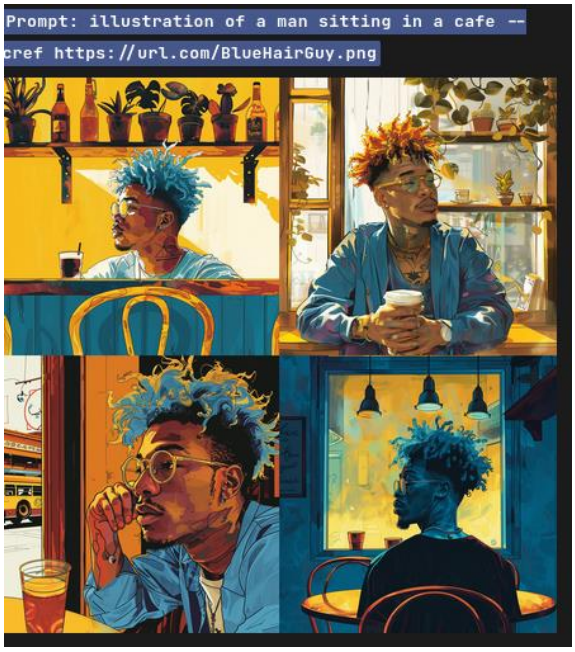

Figure 15. Images Generated by Midjourney and the Respective Prompt.

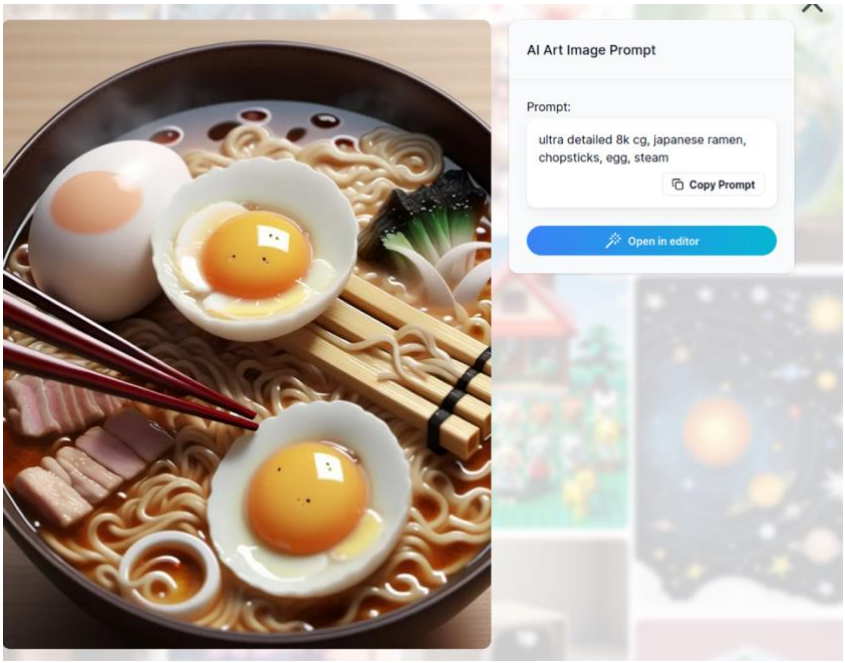

Figure 16. Image Generated by Stable Diffusion and Its Corresponding Prompt.

3.3.1.2 *Data augmentation*

Data augmentation is the process of creating new data from existing data, particularly for training new ML models. ML models require large and varied datasets for initial training, however obtaining sufficiently different real-world datasets can be difficult due to data silos, regulations, and other constraints. Data augmentation artificially expands a dataset by making minor modifications to the original data.

Model augmentation can also improve model performance. Data augmentation methods enrich datasets by generating many variations of existing data. This provides a larger dataset for training and allows a model to encounter more diverse features.

The enriched data allows the model to better generalize to unseen data and improves its overall performance in real-world contexts. In addition, data overfitting can be handled with data augmentation. It offers a significantly larger and more complete dataset for model training. It helps training sets look distinct to deep neural networks, preventing them from learning to deal with just certain features. In the following paragraphs, some popular data augmentation techniques will be described.

Image HSV (Hue, Saturation, and Value) Augmentation: This augmentation technique introduces variations in colour, lighting conditions and contrast. Alteration of hue component can simulate various lighting circumstances, such as daylight or artificial lighting, so that the model may learn to recognize objects under varying illumination levels. Adjusting the Saturation component allows us to modify the vividness or dullness of colours, exposing the model to various colour distributions and modifying the value component changes the image's brightness, allowing the model to adapt to changing brightness levels.

Image Angle/Degree rotation Augmentation: Image angle/degree augmentation includes rotating the input pictures by a certain angle or degree. Rotational changes included during training make the model more robust and capable of managing objects that occur at diverse orientations or angles in real-world photos.

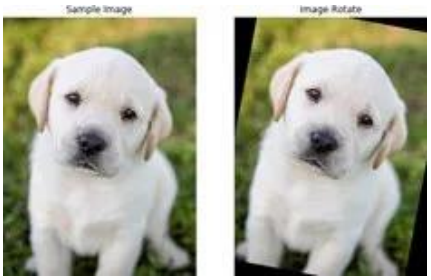

Figure 17. Example of Rotation Augmentation [106].

Image Translation Augmentation: Translation augmentation involves shifting or moving items inside an image. This approach replicates situations in which items are gently shifted or relocated inside the frame.

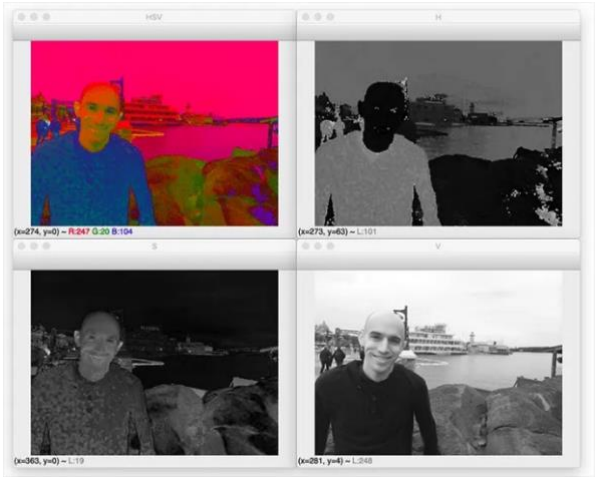

Figure 18. Example of HSVAugmentation [106].

Figure 19. An Example of Image Translation Augmentation [106].

Image Prospective Transform Augmentation: Perspective transform augmentation distorts the picture to imitate perspective shifts. This is especially beneficial in situations when items appear from multiple distances or angles.

Figure 20. An Example of Image Prospective Transform Augmentation [106].

Image Scale Augmentation: Image scale augmentation entails scaling input images to various sizes or dimensions. This augmentation trains the model to recognize things of varied sizes, allowing it to properly handle both small and large objects.

Image Shear Augmentation: Shear augmentation creates geometric deformations by tilting or skewing pictures along the x or y axes. This approach simulates real-world scenarios in which things look slanted or skewed due to perspective or camera angles.

Figure 21. An Example of Image Scale Augmentation [106].

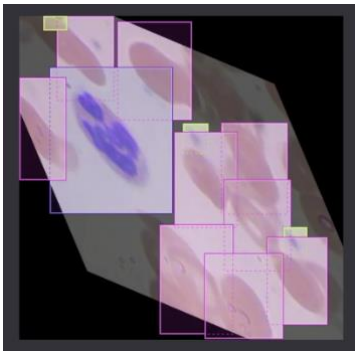

Figure 22. Example of Image Shear Augmentation [106].

Image Flip up-down(Vertically) and Flip Left-Right(Horizontally): Flip up-down augmentation includes flipping the picture vertically, creating a mirror image in which the top becomes the bottom and vice versa. Flip left-right augmentation, on the other hand, involves flipping the picture horizontally, resulting in a mirror image in which the left side becomes the right side and vice versa.

Figure 23. Example of Image Flip Up-Down (Vertically) and Flip Left-Right (Horizontally) [106].

Image Mosaic Augmentation: Mosaic augmentation is a technique for combining many pictures to generate a single training sample with a mosaic-like look. When trained on mosaic augmented pictures, the model improves its ability to handle circumstances in which objects are partially obscured or mix together. This augmentation strategy enhances the model's capacity to recognize things properly, even in difficult situations.

Image Mixup Augmentation: MixUp augmentation mixes pairs of images and their matching item labels to generate fresh training samples. This augmentation strategy increases the model's capacity to handle differences in object appearances and its overall effectiveness in finding items with comparable properties.

Figure 24. Example of Image Mosaic Augmentation [106].

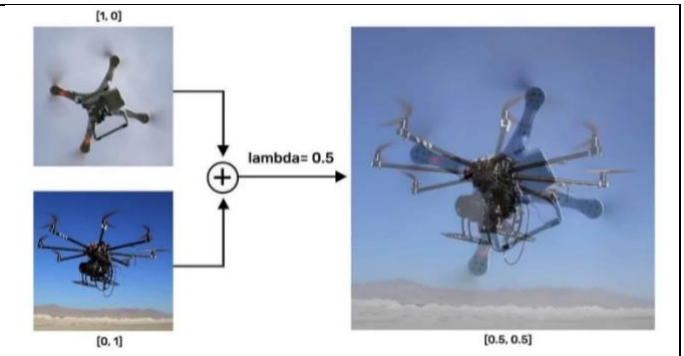

Figure 25. Example of Mixup Augmentation [106].

Image Cutmix Augmentation: CutMix augmentation involves choosing a random section of one image and pasting it onto another while keeping the relevant item labels.

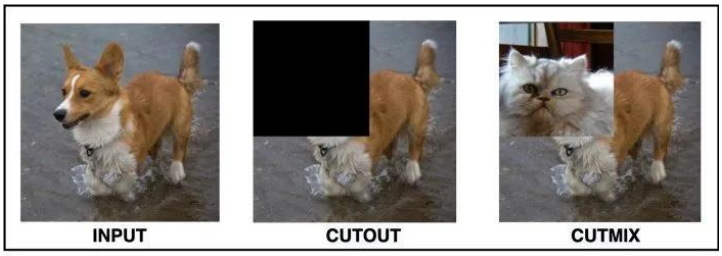

Figure 26. Cutmix Augmentation Example [106].

3.3.2 Multimedia Data Synthesis for Social Media and Web Presence

Multimedia data synthesis plays a crucial role in establishing social media and web presence. It involves high-quality multimedia content, such as images, videos, and audio files, to engage audiences and convey messages effectively. Regarding Naaman, this comprehensive process includes several steps [62] ;

1. Identify topic and application domain and use simple context-based tools to identify relevant content items.
2. Use application-specific, constrained and "knowledge-free" (unsupervised) content analysis techniques to improve precision, representation and selection of items.
3. Use the content analysis output to further improve metadata for aggregate multimedia items.
4. Leverage user interaction for improving relevance and representation.

Since presence in the web serves range of avenues to do research in the multimedia domain such as analysing community activity around multimedia resources; deriving metadata from social activity and resources; and pooling of content in social settings.

In research of Bian et al. [74] synthesis of multimedia types from social media based on three stages. The first step is *spectral filtering model* to estimate the probability that an image is relevant for a given event. Followed by a cross-media probabilistic *model* which targets discovering sub-events from microblogs of multiple media types. The last step is allowing social media documents modelling and learning the correlations between textual and visual modalities to separate the visual-representative topics and non-visual-representative topics by *multi-model event topic model*.

Various architectures have been observed in the literature, significantly contributing to multimedia data synthesis for social media. Such as, Amato et al. [83] , and Sagduyu [84], have obviously contributed to the synthesis of multimedia data for social media. Amato's research areas include multimedia summarization, which is the process of extracting and ranking important

objects from the multimedia data, using hypergraph-based modeling and influence analysis approaches. This approach is applicable in the generation of multimedia summaries and stories as has been illustrated. This method allows for the generation of concise and relevant multimedia summaries and stories, which can enhance the clarity and impact of the information presented. By effectively summarizing complex multimedia content, this approach helps in creating digestible and engaging narratives that resonate with the audience. On the contrary, Sagduyu et al.'s [84] work presents a method for creating realistic social media data with temporal features for the interactions and text contents based on topics. This system has been shown to generate data that are statistically similar to real social media data. Altogether, these works substantiate the opportunity of adopting higher-level approaches for synthesizing multimedia data for social media.

These methodologies support the need to use higher methodologies in the synthesis of multimedia data for social media. With summarization and realistic data generation with temporal features, these contributions offer solid methodologies for dealing with the vast people-generated videos on social media platforms.

Incorporating these complex models into multimedia data synthesis not only improves the timeliness, accuracy, and relevance of the synthesized material, but also ensures that the data produced accurately represents real-world phenomena. This is especially important for the use cases like the monitoring of public opinions, identifying new trends, and social media analysis where the close-to-real data is a key.

Furthermore, the constant technological developments in multimedia data synthesis (Section 3.3.1) prove that there is more to be expected in the future of this type of innovation. More advanced research could be conducted involving the use of the geospatial data as well as IoT sensor data in order to enhance the amount of multimedia content being incorporated. The use of machine learning and artificial intelligence techniques also looks promising for the automatic and improvement of the synthesis process, thus contributing to the scalability of the solutions.

To summarize, the work presented by Amato and Sagduyu shows that the field of multimedia data synthesis for social media has advanced greatly. From the above studies, it is clear that complex tools and techniques may emerge that will improve the management and use of the large and diverse multimedia data produced in today's digital environment. Therefore, if the above approaches are further developed and applied to other types of multimedia data, the results will be more effective decision making, increased user participation, and a deeper understanding of social media mechanisms in general.

3.3.3 Audio Data Synthesis

Audio data synthesis has a very long history, with the first attempts to generate speech going as far as the late 18th century [90]. In more recent times, machine learning approaches have made great progress in this domain. In 2016, WaveNet [67] was presented and is referred to as a "significant breakthrough" [88]. WaveNet is a generative, autoregressive model which utilizes dilated convolutions and in which each audio sample is conditioned on the samples of all previous timesteps [67]. Other than such autoregressive approaches, adversarial and diffusion approaches have also been used for training audio synthesis models [91]. Researchers have taken notes from the great success of large language models in (text-based) natural language processing and a new direction of textless NLP or (Generative) Speech (Large) Language Models was created. Examples of such models are AudioLM [91], VALL-E [93], and SpeechGPT-Gen [95].

AudioLM [91] is an audio-generation framework that can perform speech continuation, acoustic generation (voice conversion), unconditional generation, and even music generation (piano continuation) [122]. The authors' goal is to create a "high-quality audio generation [framework] with long-term coherent structure" [91], which they try to achieve using a combination of adversarial neural audio compression [92], self-supervised representation learning, and language modeling approaches. During training, the authors feed both acoustic (obtained using a Soundstream ([92] codec that the authors trained) and semantic (obtained using w2v-BERT XL tokens into the model. Borsos [91] then create specific experiments to assess the information that each token type carries and conclude that the acoustic tokens contain information about the speaker's identity and recording condition, while the semantic tokens contain the semantic information and attempts to generate speech without them results in lower coherence and/or inconsistencies[1] [122].

Wang [93] present VALL-E, a text-to-speech (TTS) framework that was motivated by the successes of text-based large language models and their approach to training models with an abundance of diverse data. Like such models, VALL-E shows strong in-

[1] Generation without semantic tokens.

context learning capabilities, which, for TTS models, can be defined as the ability to adapt to unseen speakers without fine-tuning [93].Furthermore, compared to the cascaded TTS systems of the time, VALL-E uses a different intermediate representation (audio codec codes instead of mel-spectrograms), defines a different training objective (language modeling instead of continuous signal regression), and is trained with around 100x more data [93]. Using a combination of autoregressive and non-autoregressive Transformer [97] decoders for the language modeling, Wang [93] achieve zero-shot TTS state-of-the-art results on two datasets. VALL-E is even capable of staying true to the emotion exhibited in the audio prompt, as well as the acoustic environment.

An open-source model similar to AudioLM and VALL-E is Suno AI's Bark [89]. Suno AI [89] state that this is a text-to-audio model, meaning it can generate not only speech but also other types of sounds when instructed to do so. Wang and Székely [94] evaluate Bark in five different domains and conclude that it produces high-quality audios but find it less robust than conventional TTS systems.

Zhang [95] present Chain-of-Information generation, a novel approach in speech generation which aims to break the process down into smaller steps and reduce the redundancies present in the current modeling approaches. Using a combination of semantic and perceptual modeling, the authors train SpeechGPT-Gen and claim that it "excels in zero-shot text-to-speech, zero-shot voice conversion, and speech-to-speech dialogue" [95] tasks. Like Wang [93], Zhang [95] also use both autoregressive and non-autoregressive models, namely SpeechGPT trained from Llama-2-7B-Chat for semantic and a Conformer for perceptual modeling, respectively. The authors conclude that their Chain-of-Information approach is one of the fastest and the fastest to converge in autoregressive and non-autoregressive modeling, respectively, when compared to two other approaches; and it also achieves the best WER and speaker similarity.

Aiming to reduce inference time and computation costs, Ye [96] propose adversarial consistency training. In this approach, pre-trained speech language models are used as discriminators, and models can be trained without needing a pre-trained diffusion teacher model [96]. The authors trained a latent consistency model [96] and named their system FlashSpeech. According to the authors, FlashSpeech's inference is around 20 times faster than the inference of other, comparable systems, while maintaining comparable performance, making their training approach a success. Aside from FlashSpeech and adversarial consistency training, Ye [96] also present a prosody generator module whose task is to improve the prosody diversity.

Siuzdak (2023) works towards bridging the gap that exists in GANs that model complex STFT coefficients and GANs that model audio in the time-domain, with the former being less successful, more challenging, as well as understudied when compared to time-domain approaches. To achieve this, the author proposes a new activation function, revisits the use of dilated convolutions, and suggests maintaining the same sampling rate throughout the modelling process. This novel approach yields Vocos, a GAN vocoder that generates STFT coefficients instead of operating in the time-domain [88]. Vocos is faster and more efficient than state-of-the-art approaches, while also matching their audio quality, and can be used for audio reconstruction, but also as a replacement vocoder for Bark [88], [89].

3.4 Gap Analysis in Synthetic Data Generation

For the gap analysis in Synthetic Data Generation, we need to consider its role as part of the CEDAR project; these techniques will be employed when necessary to complement the identified datasets, such as the ones reported in this deliverable as part of the initial Data Catalogue. There are several scenarios where synthetic data generation will be needed, as the use cases might require further data than the one identified in the data sources. Therefore, the specific requirements for synthetic data will be extracted from the identified use cases in CEDAR project, for which an initial estimation is reported in D1.1.

As it has been discussed in the previous sections, synthetic data generation techniques present well known limitations such as the difficulty to assess the characteristics of synthetic data using statistical methods, or the inheriting the bias of the original data. It could easily be argued that these are no issues, but rather a representation of the issues underlying in the original data, and as such, they would not be specific to synthetic data.

One of the main issues that lurks in the synthetic data domain, and that is pervasive to a variety of domains and techniques, resides in the fact that the internal mechanisms for the synthetic data generation are unknown. This is the problem of black box

AI, where a useful outcome has been reached but it is not clear how. While this is being addressed by explainable AI (xAI) techniques for the underlying data generation methods, it is still an open issue that could work against the main goal of generating data that is not only useful, but also worthwhile and meaningful.

Thus, it could be easy to argue that the main gap that has not been covered by synthetic data generation actually resides in the fact that the available tools and techniques have not been able to be developed in such a way that these issues have been tackled and erased from the equation. Going into detail, the generation of tabular data that lacks any bias that the original data might have had would be crucial to establish a baseline and a ground-truth to work against. This is especially crucial in situations in which the generated data does not represent an overall irregular situation, but its main purpose is to represent a regular situation. Once again, this is only an issue if the original data does only represent anomalous situations that hold no real value whenever trying to create a landscape of normality. Of course, it could easily be argued that this is an issue with the original data and not something that is actually dependent on the methods for generating synthetic data. But since synthetic data is used in multiple times where the original data is scarce or completely lacking, it should be accounted as one of the main gaps available today in the domain of synthetic data generation.

It is obvious that the main issues that the generation of multimedia synthetic data is going to face are completely different than the ones that synthetic tabular data generation faces. Although the main gap that can be detected relates to the training data, it does in a different way that it does for tabular data. A lack in diversity in the training data for synthetic multimedia data generation will lead, necessarily, to a limited generation of multimedia data. It will make the bias of the synthetic data even more apparent than for tabular data. This makes the synthetic multimedia data hard to use, as it could lead to a potential error in which the required data is not available, be it regular, be it synthetic.

Of course, any gap analysis could drag forever on nitpicking all the minor issues that have been highlighted for a set of techniques. That is, we could go over the fact that the current techniques for generating synthetic data are lacking in evaluation metrics, scalability, or conditional generation, but that would be trivial. Those are issues that, while interesting and valuable, lack the real value of being a solution that would solve the main issues that have been detected in synthetic data generation.

3.5 Methods in CEDAR for Synthetic Data Generation

CEDAR project could use synthetic data generation for all the multiple and different possibilities that have been mentioned above: the need for more data, anonymizing personal data, or creating specific data. As the initial Data Catalogue shows, the pilots have and will use tabular data, which could be complemented with the aforementioned techniques for creating synthetic tabular data. In this sense, synthetic data could be used to create a domain-agnostic use-case but using the current Data Catalogue as training data. Of course, it would also be possible to extend the number of not-so-transparent situations detected in order to empower the different analysis tools that are to be developed. And despite the availability of multiple anonymization techniques, it could be used to completely replace the personal and sensible data.

Nonetheless, all these possibilities are subsumed to the fact that it is needed. At the current stage of the project, it is unclear whether such techniques will be required to be employed within the CEDAR project. As the project activities advance It is in the future where it will be decided which are the techniques that will be used in the event that they are needed. At this point in the project is almost impossible to determine if it will be needed. Although the current Pilots and Data Catalogue can help to shed some light on what the future might bring, it is hard to know for sure. It is worth mentioning that the need for synthetic tabular data might arise in the future, since all of the data that has been collected so far is actually tabular data. With that in mind, it could be easy to imagine an extreme situation in which tabular synthetic data is needed, although in the current moment it is not. On the other hand, since none of the pilots have a clear use for multimedia data, it seems unlikely that it will be needed further down the line. Of course, if the development of the project leads to using multimedia data, the techniques listed above provide a solid foundation to develop the solution to be used.

4 Ethics Standards, Relevant Legal Framework Governmental Data

4.1 Ethics Standards in Data Solutions

Ethics, in data spaces, is very much related to the correct conduct on data sharing, by applying embedded solutions and architecture that translate European values and principles by-design. In the same sense, ethics-by-design is also a relevant point in the development of AI, which is also crucial for the CEDAR project. Additionally, considering that CEDAR is a research project, principles of research ethics must also be observed by the activities developed.

As a first ethical topic, trust is crucial for the proper development of effective data spaces, since it is an essential factor in mitigating barriers to data sharing, such as applicable regulations, fear of data breaches, and mistrust in the ownership of data [123]. Data-sharing platforms then play a relevant role in creating, maintaining and increasing trust between the different actors in a data ecosystem. So, data-sharing solutions should improve the trust of data users, guaranteeing fair access to relevant and up-to-date information and data suppliers, and ensuring that the data is secured and used for legitimate purposes.

One of CEDAR goals, is to develop and increase the trustworthiness of the solutions provided in the project, guaranteeing that good and secure practices are being put into place since the identification and preparation of datasets, until the deployment of the solutions created in the project. For this, CEDAR has foreseen a continuous and comprehensive approach for the consideration of ethical aspects since the project's beginning, what can be illustrated by the ethics workshop that took place in M3 of the project.

Additionally, the legal and ethical experts will work on transparency solutions to receive feedback on the development of the project and increase the trust of actors. Further detailed understanding of instruments tends to create more trustworthy solutions since actors have easy access to the possible (positive and negative) consequences of and understand which values and principles are considered by the developers. Thus, transparency best practices will also be installed in CEDAR design.

4.2 Human Rights

As a general ethical framework, human rights impacts shall be considered and mitigated as much as possible, since any interference to human rights should be limited to what is necessary and proportionate. Following this, different human rights instruments will be considered in CEDAR to guarantee that ethical principles are part of the project since the design of the activities and results.

Initially, as a non-exhaustive list of instruments, the following instruments will serve as a basis for the establishment of core principles, protections and rights to be observed in CEDAR: (i) the European Convention on Human Rights (ECHR), adopted by the Council of Europe (CoE) [124] in 1950 and binding to all member states, overseen by the European Court of Human Rights (ECtHR) [125]; and (ii) the European Union Charter of Fundamental Rights (the Charter), binding instrument to all Member-States of the EU, adopted in 2002 [126].

On data handling activities, the following rights must be highlighted:

- **Non-discrimination:** all individuals should be considered as equal, regardless of their characteristics; nonetheless, some groups/grounds may have special protection conceded by specific laws and considering different contexts (e.g., race, ethnicity, colour, and membership of a national minority, gender, pollical opinions), nonetheless, in principle, the rights and freedoms should be available for everyone. In CEDAR, thus, it is important to have architecture choices that guarantee that there is no by-design exclusion of certain groups or persons to participate in the results of the project;
- **Access to information**: a means of exercising other goals, such as freedom of expression, being understood as a principle of democracy [127], essential for the development of critical ideas and opinions. In the digital economy, access to information is vital for economic development. Nonetheless, the right to information has been more and more relevant for achieving social goals, including the ones foreseen at CEDAR for more transparency in public governance.
- **Privacy and data protection:** specific solutions to guarantee the operationalization of said rights are being put into place, especially in what concerns fairness in data sharing. Also, design choices play a central role in guaranteeing that the desired values are respected in the development of ethically compliant technologies [128].

4.3 Relevant Legal Framework

One of the most complex and important obstacles that data sharing initiatives face is legal constraints. Considering the possible negative effects of data handling activities, regulators worldwide have established rules that Impact data processing activities, trying to find a balance between open data initiatives and the protection of certain information, especially when personal data

is being accessed. So, regulations must evaluate the possible balance with the positive consequences that data processing might create, by evaluating the necessity and proportionality of the data processing activities.

CEDAR, as a research project with foreseen practical applications for the established use cases and beyond, has been considering these aspects since the proposal stage to guarantee that all the activities and results do consider the possible (positive and negative) impacts of the research. Since the project will handle different datasets, involving personal, anonymised, and anonymous data, the relevant regulatory framework for the project activities has been mapped and will be the compass for the development of internal guidelines and good practices to guarantee positive results of the project, including mitigation measures for the most effective use of data as presented in this Deliverable (e.g., anonymization, pseudonymization, creation of synthetic data). So, following the consistency of the relevant regulatory framework, the project will follow a risk-based approach on the developing the design of data handling actions.

A relevant aspect to be considered is the participation of non-EU partners in the consortium. So, the definition and selection of data sets will also consider the compatibility – or lack of compatibility – with third countries' norms, also considering the best international practices. In addition, to the extent that it is proven relevant for the pilots and use cases, national legislation from EU Members partners shall also be considered. Nonetheless, this more detailed analysis, not necessarily related to the initial data handling of CEDAR, will be presented and developed under the WP7 of the project.

Personal data protection is protected as a fundamental right, beyond the need for having ethics-by-design solutions embedded in the project, it is mandatory that CEDAR also adopt and consider aspects of personal data protection in the construction of the project. On this, the General Data Protection Regulation (GDPR) will be the basis for the development of the guidelines for best practices for personal data handling.

As a European Regulation, the GDPR establishes a set of principles to be observed in personal data processing activities: (i) lawfulness, fairness and transparency; (ii) purpose limitation; (iii) data minimization; (iv) accuracy; (v) storage limitation; (vi) integrity and confidentiality; and (vii) accountability. These fundamental propositions translate well European values, including the idea of self-determination of the data subject, and should be always considered as a positive direction on designing the ecosystem of data sharing to be created by CEDAR.

With similar intent, the norm also lists rights that should be guaranteed by the controllers of personal data and that may only be restricted by law when specific interests are at stake. In addition, to guarantee the protection of personal data, the Regulation follows a prohibition approach, which can be understood that no processing activity can take place unless the data subject provides their consent or there is another legitimate base for the treatment.

Finally, considering the goal of guaranteeing the free flow of data for increasing transparency in public governance and combating corruption and fraud, the rules set by the GDPR regarding international data transfers must also be considered to the extend they apply. On this topic, the architecture of CEDAR's solution must consider the participation of non-EU actors in the data space. Guidelines on the topic shall be issued considering the different stakeholders that are part of the consortium and the possibility of other third parties also accessing the CEDAR results.

Processing data activities certain purposes fall outside of the scope of the GDPR. Nonetheless, in parallel to the adoption of this regulation, another norm was adopted, the Law Enforcement Directive (LED) [129]. LED applies to the processing of personal data for the "purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security."[2]

Since CEDAR use cases may lead to activities involving prevention, investigation and detection of criminal offences, the Directive rules may also apply, especially in the final project results. On this, it is also fundamental for the development of the research to establish common grounds on the legal vocabulary to be adopted, avoiding any illegal or inaccurate identification of practices without the observation of fair trial, access to justice, right legal procedure. In addition to this observation, LED, as a Directive, must be transposed in the EU Members national legislations, which may be relevant for specific contexts in the application of CEDAR inputs. Nonetheless, the LED will be the main instrument considered for the processing of personal data for the mentioned purposes in CEDAR, to the extent where it is applicable since it is a research project, because it illustrates the main topics and points of attention.

[2] Article 1 of LED.

Regardless of the purposes related to processing personal data, privacy-by-design and by-default principles shall be observed, directly connected to the ethics-by-design approach adopted by CEDAR and the LED and the GDPR.

It is important to notice that the GDPR and the LED only apply to personal data (including pseudonymized data), but to anonymous or anonymized data (non-personal data). However, as already mentioned in the present Deliverable, CEDAR will handle mixed datasets which consist "of both personal and non-personal data. Mixed datasets represent most datasets used in the data economy and are common because of technological development such as the Internet of Thing (i.e. digitally connecting objects), artificial intelligence and technologies enabling big data analytics" [130]. So, not only will GDPR and LED apply to CEDAR, but different regulations regarding personal and non-personal data.

In the EU, effective data use and reuse has been understood as a key answer to the economic and social development by aiming at the evolution of the European Single Market and the Common European Data Space. As an initial solution for creating possibilities of exploring data, the EU adopted, in 2018, a norm to guarantee and incentivize the smooth flow of non-personal data, namely the Regulation 2018/1807 on the Free-Flow of Non-Personal Data [131], which "aims to ensure the free flow of data other than personal data within the Union by laying down rules relating to data localization requirements, the availability of data to competent authorities and the porting of data for professional users" [132].

Considering the characteristics of mixed datasets and the possible interactions between the GDPR and the Free Flow on Non-Personal Data Regulation, the European Commission issued guidelines stating that the following system for datasets containing personal and non-personal data:

- the Free Flow of Non-Personal Data Regulation applies to the non-personal data part of the dataset;
- the personal data part of the dataset must consider that the free movement of personal data within the Union shall be neither restricted nor prohibited for reasons connected with the protection of natural persons with regard to the processing of personal data[3] [133];
- when the non-personal data and the personal data parts are 'inextricably linked' (e.g., impossible to separate) the GDPR – or the LED - fully applies to the whole mixed dataset [130].

Following the understanding that we should make data accessible for improving innovation and competition, in 2019, The Directive on Open Data and Re-Use of Public Sector Information (**Open Data Directive**) [134] was adopted by the EU. The Directive aims to promote the use of open data and stimulate innovation in products and services, by establishing a set of minimum rules governing the re-use of existing documents held by public sector bodies of the Member States or by certain public undertakings, and research data [134]. For CEDAR, the provisions set by the Open Data Directive are very relevant in one hand for the inputs that the solutions will receive, which will contain various information maintained by the public sector and research, and in the other hand for guaranteeing the access of results of the research activities.

Recently, the same rationale of enhancing the use of data for economic growth, innovation and social progress was behind the creation of the Data Strategy for Data. The policy "aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty" and ensuring the human-centric development of AI by adopting legislative measures on data governance, access and reuse; making data more available; investing in data governance tools, infrastructure and mechanisms; and ensuring competitive clouding services [135]. CEDAR is, then, completely intertwined with this new European policy.

For the construction of the relevant legal framework for the project, as part of the Strategy, two Regulations should be highlighted: (i) the Data Governance Act (DGA) [136]; and (ii) the Data Act [137]. The first complements the Open Data Directive, and the last focus on the openness of private controlled data, especially industrial data [138]. The success of data spaces relies on several forms of data sharing, such as business-to-business (B2B), government-to-business (G2B), and business-to-consumer (B2C), so, CEDAR must understand and comply with the different regulatory norms, understanding the possible overlapping of the legal instruments.

Even though the Data Act mainly sets down rules for accessibility obligations to industrial data (especially regarding connected products), it still may be relevant for CEDAR, notably the provisions regarding interoperability for data spaces. Interoperability is crucial for guaranteeing the relevance of the CEDAR solutions on the Data Market and to achieve continuous and effortless data

[3] Article 1(3) of the GDPR.

flow between the instruments of the project and other data sharing tools (e.g., data spaces). Plus, industrial data may present some relevance for certain use cases and future possibilities of application of CEDAR solutions.

Crucial for the development of data spaces and other data governance initiatives, the DGA sets a system that allows a wider use of publicly or protected data in various sectors. Reinforcing the importance of the Common European Data Spaces, setting formal rules for voluntary data sharing and ensuring the availability for the flow of information, which may involve either or both private and public organisations.

Considering the CEDAR project, this Regulation is decisive for guaranteeing that the flow of data will be smooth for both the creation of the data sharing instruments (meaning the incoming data that will feed the CEDAR instruments) and also for the availability of the results of the project (meaning the outcoming data that shall be available for different actors with the development of the data sharing infrastructure created by CEDAR). Thus, it is stimulating to understand that the DGA will be a solution for the construction of CEDAR outcomes and the possible requests of data, but, at the same time, CEDAR must be compliant to the rules established by the regulation, making the data available for others.

Lastly, the DGA establishes that "the exchange of data, purely in pursuit of their public tasks, among public sector bodies in the Union or between public sector bodies in the Union and public sector bodies in third countries or international organisations, as well as the exchange of data between researchers for non-commercial scientific research purposes, should not be subject to the provisions of this Regulation concerning the re-use of certain categories of protected data held by public sector bodies"[4]. This exception already set by the Regulation, must be interpreted in parallel to the broad concept of scientific research purposes underlined by the DGA[5]. Along these lines and related to what was already presented above, it is imperative that CEDAR understands the challenges and possibilities of being a research project, but that aims to create a final product to be available beyond the life of the research activities.

Still on the public sector, the recent Interoperable Europe Act (IEA) [139] also has relevance for CEDAR activities, since aims to accelerate the digital transformation of the public sector, in better connecting public services and facilitating the international flow of data [140]. By strongly enforcing cooperation between EU public administrations, the IEA creates possibilities to share and reuse interoperability solutions via the "Interoperable Europe portal" and sets rules regarding mandatory interoperability assessments, already in consistency with the AI Act [141]. CEDAR may learn from and contribute to these initiatives, including the possibility of participating in a regulatory sandbox.

Beyond those, the project shall also evaluate the possibilities of requiring data for platforms in the scope of the Digital Services Act (DSA) [142], especially the possibilities set by Article 40 of the Regulation, which establishes a procedure for data access for researchers. In case it is understood that retrieving certain information from very large online platforms (VLOPs) or search engines (VLOSEs), CEDAR may explore the possibility of requesting access to data controlled by VLOPs or VLOSEs [143]. However, understanding the possibilities of access and use of the requested data is a pressing issue, since it is not yet completely clear the usefulness of said procedure for CEDAR and to which extend the project can exploit the data obtained via this sort of request.

Considering the need to protect the security of data, cybersecurity will also be a main concern of the project, as illustrated by the development of the cybersecurity risk assessment, part of the WP3. Nonetheless, as part of the relevant regulatory matrix for the protection of data, different pieces of legislation will also be considered in CEDAR design choices in the operations involving data: (i) NIS2 Directive; and (ii) Cybersecurity Act.

Finally, since CEDAR aims to develop and apply AI solutions in the project's activities, the new AI Act must also be considered. The risk-based approach adopted by CEDAR is aligned with the ideas behind the AI Regulation, while the possible need for specific impacts assessments will also be considered. Even though the Regulation will only be fully applicable 2 years after the publication in the official Journal, the legal provisions will already be taken into consideration by CEDAR to guarantee that the project is compliant with the obligations and best practices set by the act, especially if any AI solution used in the project may fall under the definition of AI systems or the prohibited practices of AI (e.g., use of biometric identification systems by law enforcement).

[4] Recital 12 of the DGA.
[5] DGA establishes scientific research purposes as "any type of research-related purpose regardless of the organization or financial structure of the research institution in question, with the exception of research that is being conducted by an undertaking with the aim of developing, enhancing or optimizing products or services".

Table 1. Regulatory Matrix for CEDAR.

Regulatory matrix for CEDAR	
Norm	Observations
European Convention on Human Rights	Binding instrument for State Members – including Ukraine - of the Council of Europe, establishing minimum human rights that must be observed and shall be overseen by the European Court of Human Rights.
European Union Charter of Fundamental Rights	With very similar provisions to the European Convention on Human Rights, this binding document to EU Members also lays down a set of human rights which should be considered by different activities, and that translate the values and principles
General Data Protection Regulation	Lays down rules relating to the protection of personal data and to the free movement of personal data, by protecting the fundamental rights and freedoms of natural persons, establishing a set of principles and rights to be observed in data processing activities. It is highlighted that the free movement of personal data within the EU shall be neither restricted nor prohibited by the protection of personal data, but specific rules shall be observed for the international transfer of personal data.
Law Enforcement Directive	Applies for the processing of personal data for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security. Considering the use cases and the goals of CEDAR, LED may apply and shall be considered. Nonetheless, the possible evaluation of the LED must also acknowledge that CEDAR is a research project that may lead to available final projects.
Free Flow of Non-Personal Data Regulation	Aims to ensure the free flow of non-personal data in the EU, relaxing rules regarding the localization requirements, the availability of data to competent authorities and the porting of data for professional users.
Open Data Directive	Lays downs rules to guarantee the openness of data held by the public sector and research data.
Data Governance Act	Expands the possibility of use and reuse of data – both personal and non-personal -, while creating mechanisms that would allow an effective exploration of data, while observing the European values, principles and rights.
Data Act	Establishes rules on different types of relationships of data sharing and on access rights, while also incentivizing interoperability between data spaces.
Interoperable Europe Act	With the goal of providing tools for more efficient public services across the EU, the Regulation creates an ecosystem of shared interoperability solutions, including regulatory sandboxes.
Digital Services Act	Aims to contribute to the proper functioning of the internal market for intermediary services by setting out harmonized rules for a safe, predictable and trusted online environment, facilitating innovation and observing fundamental rights. For CEDAR, the possibility of data access for researchers (article 40) may gain relevance.
NIS2 Directive	By establishing different rules for different categories of entities, the Directive aims to increase the cybersecurity standards in the EU. Even if not directly applied to CEDAR, it may present interesting guidelines for ensuring good practices on cybersecurity.
Cybersecurity Act	General Act for the goals of more homogeneous and relevant cybersecurity approaches, while reaffirming the importance of cybersecurity in the EU.
AI Act	General Regulation on AI, explaining high-risk AI categories, rights of consumers and banned uses of these technologies.

4.4 Data Solutions

The vast amount of data being collected today is essential for analysis, research, and strategic development but simultaneously poses substantial risks to individual privacy. In response to these concerns, the General Data Protection Regulation was established to unify data privacy laws across Europe, setting a global benchmark for data protection. These regulations mandate that entities implement technical and organizational measures to ensure data confidentiality and protect personal information. Privacy-preserving techniques are crucial in this context as they aim to prevent unauthorized data access and de-identify personal information, balancing the need for privacy with the requirements of data utility. Although achieving this balance is challenging, it is essential to safeguard individuals' privacy without significantly compromising the effectiveness of data-driven tasks.

Different types of data require specific privacy-preserving techniques due to their unique characteristics. We can distinguish three main data modalities: structured, semi-structured, and unstructured data. Structured data has a defined model, format, and structure, where techniques like encryption and anonymization can be effectively applied. Semi-structured data has an apparent pattern, possessing some organizational properties but still allowing for variability, where schema-based anonymization techniques are often used. Unstructured data lacks a predefined format, requiring methods like natural language processing for redacting sensitive information. Some examples for each data modality are presented in Figure 27.

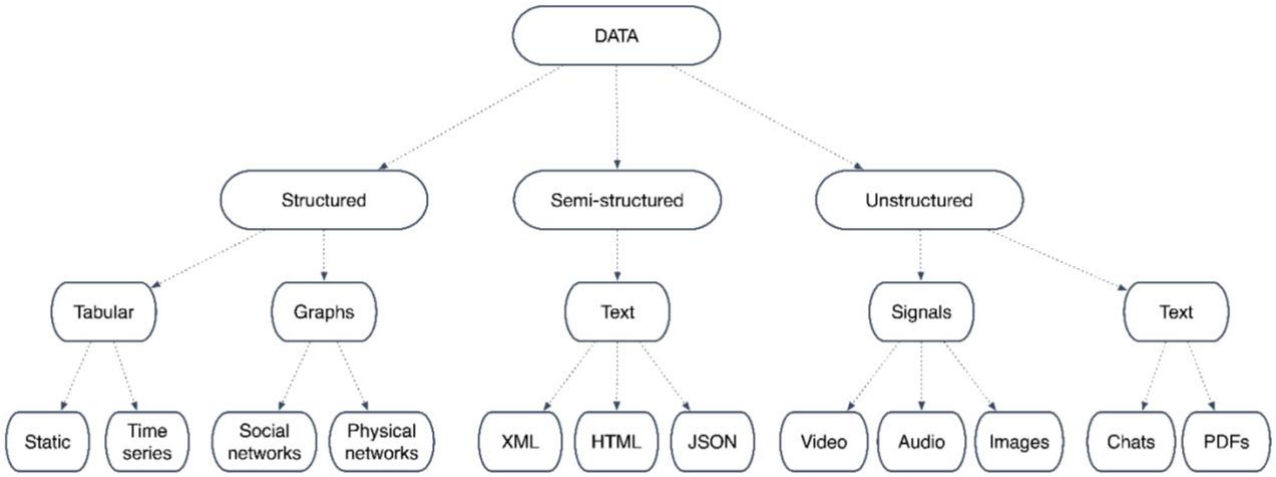

Figure 27. Taxonomy of Data Modalities.

4.4.1 Privacy Risks for Textual Data

Three types of privacy threats are commonly considered when anonymizing structured and semi-structured textual data:

1. **Membership disclosure:** an attacker can determine if data about a certain individual is present in the dataset.
2. **Attribute disclosure**: an attacker can determine new characteristics of a certain individual based on the information available in the released data. As an example, linkage to a set of data entries allows inferring information if all items share a certain sensitive attribute value.
3. **Identity disclosure**: an attacker can link a record in the released data to a specific individual. This is a serious type of attack, as it has legal consequences for data owners according to many laws and regulations worldwide.

A data controller can validate the effectiveness of a privacy-preserving technique through privacy measures appropriated for each privacy threat. A review of existing risk disclosure measures is presented in Table 2.

Table 2. Summary of Risk Disclosure Measures for Semi-Structured and Structured Textual Data.

Disclosure Risk	Risk Measure Type	Risk Measure Name	Attribute
Identity	Singling out	*k*-anonymity [1]	Categorical
		k-map [2]	
	Probabilistic modeling	Poisson-log-normal [3]	
		Logarithmic series [4]	
		Pitman [4]	
		Gaussian copulas [5]	
	Special uniques	SUDA2 [6]	
	Outliers	Standard deviation-based intervals [7]	Numerical
		Distance-based intervals [8]	
	Clustering	Density-based [9]	
Identity/ attribute	Record linkage	Distance-based [10]	Numerical and Categorical
		Rank-based [11]	
	k-anonymity-based	(α, k)-anonymity [12]	Categorical

		l-diversity [13]	
		(p, k)-angelisation [14]	
Attribute	*k*-anonymity-based	(X, Y)-anonymity [15]	Categorical
		t-closeness [16]	
		σ-disclosure privacy [17]	
		β-likeness [18]	

4.4.2 State of the Art in Anonymization of Textual Data

The goal of privacy-preserving techniques (PPTs) is to release a modified dataset that reduces the disclosure risk while still allowing us to perform statistical analysis and data mining tasks. PPTs for semi-structured and structured textual data can be categorized into four groups:

1. **Non-perturbative (NP)**: their goal is to reduce the amount of information in the data by reducing the level of detail or partially suppressing information while preserving the truthfulness.
2. **Perturbative (P)**: they distort data while guaranteeing that the statistics computed on the perturbed data do not differ significantly from statistics obtained on the original data set.
3. **De-associative (DA)**: they create buckets to break the correlation between the sensitive and the other attributes.
4. **Pseudonymization (PS)**: they replace sensitive attribute values with cryptographically generated tokens, hence minimizing the exposure of sensitive data while preserving data utility and accuracy. Tokens can be reversible or irreversible.

Table 3 summarizes the techniques considered in CEDAR and detailed in the following subsections.

Table 3. Summary of Privacy-Preserving Techniques for Semi-Structured and Structured Textual Data.

Type	Technique Name	Attribute	Principle
NP	Global recoding	Numerical and Categorical	Combines several categories to form more general categories
	Local recoding [19]		Recodes into broader categories only some values
	Top-and-bottom coding		Replaces values above or below a defined threshold
	Suppression		Deletes or replaces cells/rows/attributes
	Sampling		Selects a sample of the original microdata
P	Swapping [[20]-[21]]		Exchanges values of certain attributes across records
	Noise [[22]-[23]]	Numerical	Adds/subtracts/multiplies random values
	Microaggregation [[24],[25],[26],[27],[28],[29]]	Numerical and Categorical	Groups similar values and assigns an aggregated value to the group
	Rounding	Numerical	Replaces values with rounded ones
	PRAM	Categorical	Reclassifies the values according to the Markov matrix
	Shuffling [30]	Numerical	Uses a regression model to determine which values to exchange
DA	Bucketisation [[31]-[32]]	Categorical	Bucketises data and permutes sensitive values
	Anatomisation [[33],[34],[35]]		Bucketises data and publishes a QIT and a ST
	Angelisation [36]		Divides data into batches and then buckets
	Slicing [[37],[38],[39]]		Partitions data vertically and horizontally
PS	Deterministic encryption		Replaces values with one-way to two-way tokens
	Format-preserving encryption		Replaces values with one-way to two-way format-preserving tokens
	Hashing		Replaces values with encrypted hashed one-way tokens

4.4.2.1 Non-perturbative Techniques

Global recoding (a.k.a. generalization) combines several categories to create more general categories. The application of global recoding in a categorical attribute *V* results in a new V' with a lower number of possible values. For a continuous attribute, V is

replaced by a discretized version of V. The main objective of global recoding is to divide the tuples in the dataset into a set of disjoint equivalence classes and then transform the attribute values of the tuples in each equivalence class to the same format.

Local recoding recodes into broader categories when necessary. The replacement can be partial, i.e., only some occurrences of the attribute *V* are replaced. Therefore, it can have a lower cost in terms of information loss.

Top-and-bottom coding is a special case of recoding applicable to continuous or categorical ordinal attributes. A top recoding covers values of an attribute above a specified upper threshold, while a bottom coding covers values below another threshold.

Suppression suppresses data. There are three common levels of suppression:

1. *Cell suppression* replaces infrequent values of an attribute in a record with a special character. Cell suppression significantly reduces predictive performance.
2. *Tuple suppression* hides whole tuples from the released data. This technique may disturb the truthfulness of the released data.
3. *Attribute suppression* hides an attribute from the released data. This approach is useful when a categorical attribute has many distinct values or when two attributes are highly correlated.

Sampling releases a sample of the original dataset. It is suitable for categorical attributes. However, it may be inappropriate for continuous attributes due to the presence of many distinct values.

4.4.2.2 Perturbative Techniques

Swapping techniques can be divided into two groups: data swapping and rank swapping.
Data swapping exchanges values of certain attributes across records. Rank swapping ranks the values of the attribute in ascending order; then it swaps each ranked value with another ranked value randomly chosen within a restricted range. The range typically corresponds to the number of swapped values.
Data swapping has many advantages, namely: *(i)* it removes the relationship between the record and the individual; *(ii)* it can be used in one or more sensitive attributes without disturbing the non-sensitive attributes; *(iii)* no non-sensitive attributes are deleted; *(iv)* it provides protection to the rare and unique values; and *(v)* it is not limited to the type of attributes.
This technique has also some drawbacks, namely *(i)* it can produce many records with unusual combinations; *(ii)* it can severely distort the statistics on any sub-domain; and *(iii)* it does not prevent attribute disclosure, as it only reorders data.

Noise techniques protect personal data via additive or multiplicative noise.
Three main procedures have been developed for additive noise:

1. Uncorrelated noise addition replaces the vector of observations x_j or the j-th attribute in the original dataset with the vector $z_j = x_j + \varepsilon_j$, where ε_j denotes normally distributed errors derived from a Normal distribution $N(0, \sigma^2_{\varepsilon j})$ (white noise).
2. Correlated noise addition generates an error matrix ε^* under the restriction $\Sigma^* = \alpha\Sigma$ (correlated noise). All elements of the covariance matrix of the perturbed data diverge from those of the original data by a factor $1+\alpha$. Using correlated noise addition produces a data set with higher analytical validity than uncorrelated noise addition.
3. Noise addition and linear transformation ensures that the sample covariance matrix of the changed attributes is an unbiased estimator for the covariance matrix of the original attributes. This technique uses an additive noise on the original attributes with covariances of the errors proportional to those of the original attributes. This approach cannot be applied to discrete attributes as it does not preserve the univariate distributions of the original data.

Multiplicative noise avoids that small values are strongly perturbed, and large values are weakly perturbed.
Let X be the original numerical data and W be the perturbation attributes with expectation 1 and variance $\sigma^2_w > 0$. The perturbed data X^α is obtained by $X^\alpha = W \odot X$.

In **microaggregation**, the records in the dataset are partitioned into *g* groups of *k* or more individuals. Each value of a continuous attribute in each record is replaced by an aggregate value of the group to which the record belongs. Groups can be of variable size and may have maximal homogeneity.
Microaggregation involves three main criteria: how the homogeneity of groups is defined, the clustering algorithms used to find the homogeneous groups, and the aggregated function to replace the attribute values.

Rounding replaces the original values of the attributes with rounded values. For a given attribute, rounded values are chosen among a set of rounding points defining a rounding set.

Post RAndomisation Method (PRAM) re-codes the values of one or more categorical attributes with a certain probability, independently of each record. Let ξ be a categorical attribute in the original dataset and X be the same attribute in the perturbed

set. The probability that an original score $\xi = k$ is re-coded with the score $X = l$ is $P(X = l \mid \xi = k)$. These probabilities must be chosen appropriately to avoid unlikely combinations.

Shuffling replaces sensitive attributes by generating new data with similar distributional properties. Let X be the sensitive attributes and S be the non-sensitive attributes. It generates new data Y using the conditional distribution of X given S. The generated values are also ranked, and each X value is replaced with another X value with the rank that corresponds to the rank of the Y value.

4.4.2.3 De-associative Techniques

Bucketisation creates buckets in such a way that each record in the bucketised set corresponds to multiple sensitive values. The bucketised data consists of a set of buckets with permuted sensitive values.

Anatomy follows the bucketisation principles but, instead of permuting sensitive values, aims to publish two separate tables: a QI table (QIT) and a sensitive table (ST). Given a set V of quasi-identifiers (QI), i.e., attributes that when combined can uniquely identify an individual, a categorical sensitive attribute S, and an equivalence class with m QI-groups, this technique produces a QIT in the form of (V_1, V_2, ..., V_i, Group-ID) and a ST in the following format (Group-ID, S, Count). Each QI-group involves at least l tuples. For each QI-group and each distinct S value v in QI_j ($1 \leq j \leq m$), the ST has a record of the form: (j, v, $c_j(v)$), where $c_j(v)$ is the number of tuples in QI_j with attribute value v. This approach releases the QI and sensitive attributes in two separate sets conserving the unique common attribute, the Group-ID. Hence, when an intruder tries to join the two tables, he will not be able to associate the sensitive value with the right individual.

Angelisation starts by dividing the data into batches B_1, B_2, ..., B_b, where each batch is a set of tuples, and the sensitive attribute distribution in each batch satisfies a certain objective de-identification principle. Then, it creates another partition but into buckets, C_1, C_2, ..., C_e where each bucket is a set of at least k tuples. Given a batch and bucket partitions, an angelisation corresponds to a pair of a batch table (BT) ⟨Batch-ID, S, Count⟩ and generalized table (GT) that has all QI attributes with the column of Batch-ID where all tuples in the same bucket have equivalent generalized QI values.

Slicing groups several QIs with the sensitive attribute, preserving attribute correlations. The intuition of slicing is the partition of a dataset vertically and horizontally. Vertical partitioning groups attributes into columns based on the correlation of the attributes. Horizontal partitioning groups tuples into buckets. For each bucket, values in each column are randomly permuted to remove the linking between different columns.

4.4.2.4 Pseudonymization Techniques

In **deterministic encryption,** an input value is replaced with a value encrypted using an encryption algorithm with a cryptographic key. This method produces a hashed value, so it does not preserve the character set or the length of the input value. Encrypted, hashed values can be re-identified using the original cryptographic key and the entire output value.

In **format-preserving encryption,** an input value is replaced with a value encrypted using a format-preserving encryption algorithm that preserves both the character set and the length of the input value. Encrypted values can be re-identified using the original cryptographic key and the entire output value.

In **cryptographic hashing,** an input value is replaced with a value encrypted and hashed. The hashed output of the transformation is always the same length and cannot be re-identified.

4.4.3 State of the Art in Unstructured Data Anonymization

Unstructured data such as images, videos, and text can contain a considerable amount of personal sensitive information such as human faces, vehicle license plates, voiceprints, and personal health records.

A body of research aims to provide privacy guarantees by (i) representing the unstructured data with vectors (embeddings), (ii) obfuscating the vectors by adding well-calibrated noise, and (iii) reverting the vectors to obtain obfuscated unstructured data.

Table 4 summarizes relevant differentially private mechanisms for unstructured data content. We restrict our literature review to mechanisms for images and unstructured text.

Table 4. Differentially Private Mechanisms for Unstructured Data.

Private Data	Vectorization	Privacy Model	Privacy Mechanism
Human Face	Pixelization	Pixel DP for aggregated pixels [40]	Laplace

	GAN Latent Coding	Latent Vector DP [41]	Laplace
	SVD	Euclidean Privacy for Individual Images [42]	Multivariant Laplace
Sensitive Text	GloVe, FastText	Euclidean Privacy for Individual Words [43]	Multivariate Normal Gamma Distribution
	Pretrained Word Embedding and BERT	Euclidean Privacy and LDP for Sensitive Tokens [44]	Exponential Mechanism Mangat's Randomized Response

4.4.3.1 *Differentially private Methods for Image Content*

The most popular methods to obfuscate Regions of Interest (ROIs) in images can be categorized into *pixelization*, *blacking*, and *blurring*.

For example, DartBlur [45] uses a DNN architecture to generate blurred faces that effectively hide facial privacy information; Disguise [46] de-identifies facial images by extracting and substituting depicted identities with synthetic ones, generated using variational mechanisms to maximize obfuscation and non-invertibility; and [47] proposes a head inpainting obfuscation technique that generates realistic head in paintings in social media photos.

Pixel DP. This work adapts the differential privacy model to the image domain, by pixelizing the input image as grid cells before applying a Laplace perturbation. The noise scale is controlled by a sensitivity parameter calculated from the grid cell size, the number of differing pixels, and the privacy budget.

Latent Vector DP (LDP). This work provides DP guarantees for image latent vector representations. Privacy budgets are allocated to various elements in the latent space according to their weights, and calibrated Laplace noise is added to images in the semantic space before GAN is used to synthesize realistic-looking faces.

Euclidean Privacy. This work generates privatized realistic-looking faces by exploiting GANs to learn image semantics such as smiling and young attributes to obtain latent codes for facial images. The perception indistinguishability notion is then defined for latent codes' representation of images, while perception distance between latent codes is defined as the semantic dissimilarity between images.

4.4.3.2 *Differentially private Methods for Text Content*

Several works have proposed DP mechanisms by tailoring distance metrics. These works use word embedding models such as GloVe, FastText, word2vec, and Hyperbolic embedding, to transform textual data into high-dimensional vectors; then, they perturb the embeddings with noise vectors sampled from probabilistic distributions; and finally, they project back the noisy real-numbered vectors to their closest words in the embedding model.

Euclidean Privacy. [48] achieves word's privacy by multivariate normal distribution and gamma distribution. The similarity of word pairs is defined as the accumulated sum of individual word pairs' Euclidean distance. The magnitude of the noise vector is sampled from a gamma distribution, while the weight of each element in the vector is sampled with a multivariate normal distribution.

[49] divides tokens including characters and word pieces into sensitive and non-sensitive types and then only sanitizes sensitive ones to improve overall utility. The definition of sensitive tokens varies across individuals. A personalized mechanism can be introduced to tag sensitive tokens on users' sides. The work uses the Utility Optimized Metric LDP, which considers both sensitivity and similarity levels between input pairs.

4.4.3.3 *Utility and Privacy Analysis for Unstructured Data*

While privacy loss is quantified with the ε parameter, utility loss is measured between the real data and its obfuscated version in the experimental evaluations. Different utility loss metrics are adopted to quantify utility losses in different data types.

Utility Analysis for Images. Several utility metrics are adopted to quantify the difference between the real image and its obfuscation, including MSE; structural similarity index (SSIM), which compares luminance, contrast, and structure; and Fréchet Inception Distance (FID), which is used to measure image quality in GANs.

Utility Analysis for Text. Utility metrics used to quantify the loss introduced by perturbing the embedding vectors of words are the F1-score, number of correct predictions, accuracy, mean average precision, AUC, and mean reciprocal rank.

These metrics are used to evaluate accuracy for topic and text classification tasks.

4.4.4 Workplan in CEDAR for Multimodal Data (Pseudo)Anonymization

To evaluate the effectiveness of privacy-preserving techniques for structured and semi-structured textual data, we propose a comprehensive evaluation pipeline. The first step involves data preprocessing, where attributes are categorized into four distinct categories: identifiers (I), quasi-identifiers (QI), sensitive (S), and non-sensitive. Identifiers, such as names and social security numbers, directly identify individuals and should be either removed or replaced by pseudonyms. Quasi-identifiers, like date of birth, gender, location, and profession, can uniquely identify individuals when combined. Sensitive attributes include highly critical information like religion, sexual orientation, and political opinions, while non-sensitive attributes do not contain sensitive information.

The second step is the application of various privacy-preserving techniques identified in our preliminary analysis, which include global recoding, local recoding, top-and-bottom coding, noise addition, microaggregation, PRAM, angelisation, and slicing. Each technique will be applied to generate perturbed versions of the dataset.

The perturbed datasets will be evaluated along two verticals: disclosure risk and predictive performance. Risk assessment will be conducted using several representative risk disclosure measures, such as k-anonymity, l-diversity, t-closeness, β-likeness, and distance-based intervals.

For datasets intended for data mining and machine learning tasks, the utility of the perturbed data will be evaluated in terms of data mining/ machine learning workloads. Predictive models will be built from the perturbed data, and the prediction accuracy will be used to assess the usefulness of these datasets. Typical measures for assessing predictive performance in classification include Precision and Recall, (balanced) Accuracy, (weighted) F-score, Geometric Mean, and (weighted) AUC. Classification algorithms that will be considered include Random Forest, Bagging, XGBoost, and Support Vector Machine. For regression tasks, measures such as Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error will be used, with algorithms like linear regression, logistic regression, and regression trees. Clustering analysis will be evaluated using measures like the Rand index, Davies-Bouldin index, Fowlkes-Mallow's index, and Silhouette score.

The whole pipeline is illustrated in Figure 28.

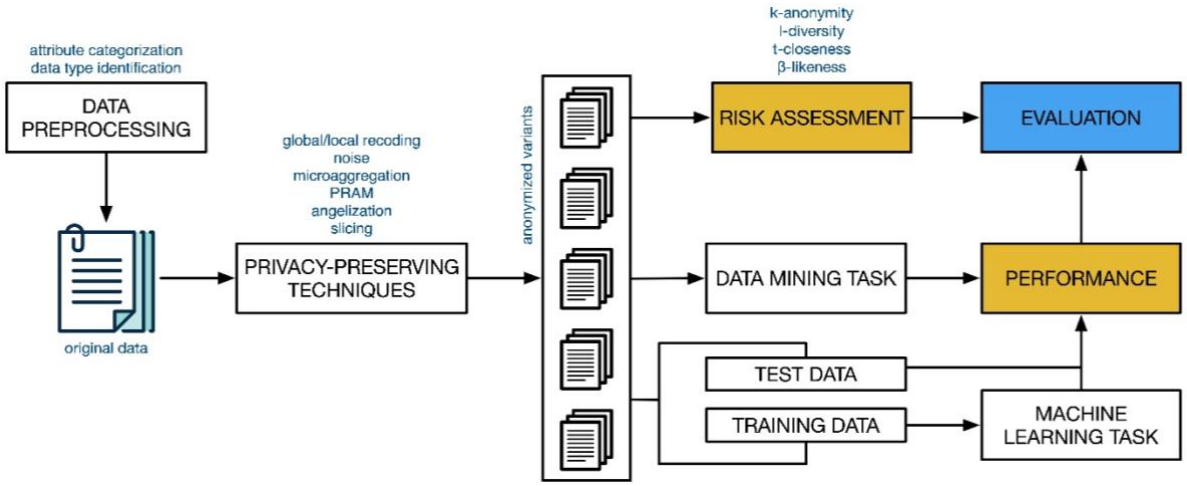

Figure 28. Pipeline for Evaluating Privacy-Preserving Techniques for Structured and Semi-Structural Textual Data.

A similar pipeline will be defined to evaluate privacy-preserving techniques for unstructured data.

5 Conclusion

Since it is reported in this deliverable, CEDAR makes significant progress towards the development of high-quality, analytics-ready and open data to increase transparency in public governance. Until the M6, the project was able to identify and acquire multiple data source types, including pilot partners' internal databases, data accessible through APIs, and extracted data from the web. Meanwhile the gaps of data collection were identified, new synthetic data generation approaches were used to ensure data was enriched and complex as real data.

One of the most important results of this phase is the understanding of the Initial Data Catalogue, which combines real data sets with synthetic data that is missing or fills the gaps in the data sets. Furthermore, ethical and legal issues that may have effect on the project are well addressed in Section 4. By this way, anonymization techniques are used, and all data processing processes are brought into compliance with certain laws to protect the security and confidentiality of data. This ethical framework aims to protect the individual privacy meanwhile supports the measurability, accuracy, and reliability of data collected and shared.

Based on the findings and approaches determined in this first stage, the Data Catalogue will be gradually developed and at the same time, continuous control of the results obtained will be conducted. Later phases of the project will focus more on improving the quality and usefulness of the data, as well as developing better methods for updating data sources and synthesizing synthetic data. These efforts will be vital to confront new problems and opportunities in public administration based on the management of big data.

The results of the first phase of CEDAR therefore confirm the effectiveness of the project in revolutionizing public administration by providing complex, open and voluntary ethical data. On this solid foundation, CEDAR is ready to further its goals and make a major contribution to the development of public administration analytics.

6 List of References

[1] P. Samarati. Protecting respondents identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.

[2] K. El Emam and F. K. Dankar. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association, 15(5):627–637, 2008.

[3] C. J. Skinner and M. Elliot. A measure of disclosure risk for microdata. Journal of the Royal Statistical Society: series B (statistical methodology), 64(4):855–867, 2002.

[4] N. Hoshino. Applying pitman's sampling formula to microdata disclosure risk assessment. Journal of Official Statistics, 17(4):499, 2001.

[5] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1):1–9, 2019.

[6] A. M. Manning, D. J. Haglin, and J. A. Keane. A recursive search algorithm for statistical disclosure assessment. Data Mining and Knowledge Discovery, 16(2):165–196, 2008.

[7] T. M. Truta, F. Fotouhi, and D. Barth-Jones. Global disclosure risk for microdata with continuous attributes. In Privacy and Technologies of Identity, pages 349–363. Springer, 2006.

[8] M. Templ and B. Meindl. Robust statistics meets sdc: New disclosure risk measures for continuous microdata masking. In International Conference on Privacy in Statistical Databases, pages 177– 189. Springer, 2008.

[9] D. Ichim. Disclosure control of business microdata: A density-based approach. International statistical review, 77(2):196–211, 2009.

[10] V. Torra, J. M. Abowd, and J. Domingo-Ferrer. Using mahalanobis distance-based record linkage for disclosure risk assessment. In International Conference on Privacy in Statistical Databases, pages 233–242. Springer, 2006.

[11] K. Muralidhar and J. Domingo-Ferrer. Rank-based record linkage for re-identification risk assessment. In International Conference on Privacy in Statistical Databases, pages 225–236. Springer, 2016.

[12] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 754–759, 2006.

[13] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):3–es, 2007.

[14] A. Anjum, N. Ahmad, S. U. Malik, S. Zubair, and B. Shahzad. An efficient approach for publishing microdata for multiple sensitive attributes. The Journal of Supercomputing, 74(10):5127–5155, 2018.

[15] K. Wang and B. C. Fung. Anonymizing sequential releases. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 414–423, 2006.

[16] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 IEEE 23rd International Conference on Data Engineering, pages 106–115. IEEE, 2007.

[17] J. Brickell and V. Shmatikov. The cost of privacy: destruction of data-mining utility in anonymized data publishing. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 70–78, 2008.

[18] J. Cao and P. Karras. Publishing microdata with a robust privacy guarantee. 2012.

[19] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 785–790, 2006.

[20] J. Domingo-Ferrer and V. Torra. Disclosure control methods and information loss for microdata. Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies, pages 91–110, 2001.

[21] J. Nin, J. Herranz, and V. Torra. Rethinking rank swapping to decrease disclosure risk. Data & Knowledge Engineering, 64(1):346–364, 2008.

[22] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez. Enhancing data utility in differential privacy via microaggregation-based k-anonymity. The VLDB Journal, 23(5):771–794, 2014.

[23] C. Dwork. Differential privacy. In International Colloquium on Automata, Languages, and Programming, pages 1–12. Springer, 2006.

[24] J. Domingo-Ferrer, A. Martinez-Balleste, J. M. Mateo-Sanz, and F. Sebe. Efficient multivariate data-oriented microaggregation. The VLDB Journal, 15(4):355–369, 2006.

[25] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and data Engineering, 14(1):189–201, 2002.

[26] M. Laszlo and S. Mukherjee. Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering, 17(7):902–911, 2005.
[27] S. Martinez, D. Sanchez, and A. Valls. Semantic adaptive microaggregation of categorical microdata. Computers & Security, 31(5):653–672, 2012.
[28] V. Torra. Microaggregation for categorical variables: a median based approach. In International Workshop on Privacy in Statistical Databases, pages 162–174. Springer, 2004.
[29] K. Muralidhar and R. Sarathy. Data shuffling—a new masking approach for numerical data. Management Science, 52(5):658–670, 2006.
[30] B. Li, Y. Liu, X. Han, and J. Zhang. Cross-bucket generalization for information and privacy preservation. IEEE Transactions on Knowledge and Data Engineering, 30(3):449–459, 2017.
[31] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In 2007 IEEE 23rd International Conference on Data Engineering, pages 116–125, 2007.
[32] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In Proceedings of the 32nd international conference on Very large data bases, pages 139–150, 2006.
[33] K. Zhiwei, W. Weimin, Y. Shuo, F. Hua, and Z. Yan. Research progress of anonymous data release. In 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), pages 226–230. IEEE, 2017.
[34] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang, and B. Shi. Permutation anonymization: Improving anatomy for privacy preservation in data publication. In L. Cao, J. Z. Huang, J. Bailey, Y. S. Koh, and J. Luo, editors, New Frontiers in Applied Data Mining, pages 111–123, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
[35] Y. Tao, H. Chen, X. Xiao, S. Zhou, and D. Zhang. Angel: Enhancing the utility of generalization for privacy preserving publication. IEEE transactions on knowledge and data engineering, 21(7):1073–1087, 2009.
[36] T. Li, N. Li, J. Zhang, and I. Molloy. Slicing: A new approach for privacy preserving data publishing. IEEE transactions on knowledge and data engineering, 24(3):561–574, 2010.
[37] J. Han, F. Luo, J. Lu, and H. Peng. Sloms: A privacy preserving data publishing method for multiple sensitive attributes microdata. J. Softw., 8(12):3096–3104, 2013.
[38] E. K. Budiardjo, W. C. Wibowo, et al. Privacy preserving data publishing with multiple sensitive attributes based on overlapped slicing. Information, 10(12):362, 2019.
[39] L. Fan. Image pixelization with differential privacy. In Data and Applications Security and Privacy XXXII: 32nd Annual IFIP WG 11.3 Conference, DBSec 2018, Bergamo, Italy, July 16–18, 2018, Proceedings 32, pages 148–162. Springer, 2018.
[40] T. Li and C. Clifton. Differentially private imaging via latent space manipulation. arXiv preprint arXiv:2103.05472, 2021.
[41] J.-W. Chen, L.-J. Chen, C.-M. Yu, and C.-S. Lu. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6478–6487, 2021.
[42] O. Feyisetan, B. Balle, T. Drake, and T. Diethe. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In Proceedings of the 13th international conference on web search and data mining, pages 178–186, 2020.
[43] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. Chow. Differential privacy for text analytics via natural text sanitization. arXiv preprint arXiv:2106.01221, 2021.
[44] B. Jiang, B. Bai, H. Lin, Y. Wang, Y. Guo, and L. Fang. Dartblur: Privacy preservation with detection artifact suppression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16479–16488, 2023.
[45] Z. Cai, Z. Gao, B. Planche, M. Zheng, T. Chen, M. S. Asif, and Z. Wu. Disguise without disruption: Utility-preserving face de-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 918–926, 2024.
[46] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5050–5059, 2018.
[47] O. Feyisetan, B. Balle, T. Drake, and T. Diethe. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In Proceedings of the 13th international conference on web search and data mining, pages 178–186, 2020.
[48] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. Chow. Differential privacy for text analytics via natural text sanitization. arXiv preprint arXiv:2106.01221, 2021.
[49] Stenger, Michael & Leppich, Robert & Foster, Ian & Kounev, Samuel & Bauer, André. (2023). Evaluation is Key: A Survey on Evaluation Measures for Synthetic Time Series. 10.21203/rs.3.rs-3331381/v1.
[50] Isasa, I., Hernandez, M., Epelde, G. et al. Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. BMC Med Inform Decis Mak 24, 27 (2024). https://doi.org/10.1186/s12911-024-02427-0

[51] Stenger, Michael & Leppich, Robert & Foster, Ian & Kounev, Samuel & Bauer, André. (2023). Evaluation is Key: A Survey on Evaluation Measures for Synthetic Time Series. 10.21203/rs.3.rs-3331381/v1.

[52] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401-4410).

[53] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, *34*, 852-863.

[54] de Souza, V. L. T., Marques, B. A. D., Batagelo, H. C., & Gois, J. P. (2023). A review on generative adversarial networks for image generation. *Computers & Graphics*.

[55] Gao, H., Pei, J., & Huang, H. (2019, July). Progan: Network embedding via proximity generative adversarial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1308-1316).

[56] Viazovetskyi, Y., Ivashkin, V., & Kashin, E. (2020). Stylegan2 distillation for feed-forward image manipulation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16* (pp. 170-186). Springer International Publishing.

[57] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in neural information processing systems*, *33*, 12104-12114.

[58] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, *34*, 852-863.

[59] Wang, Z., She, Q., & Ward, T. E. (2021). Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, *54*(2), 1-38.

[60] Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019, May). Self-attention generative adversarial networks. In *International conference on machine learning* (pp. 7354-7363). PMLR.

[61] Brock, A., Donahue, J., & Simonyan, K. (2018, September). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

[62] Naaman, M. (2010). Social multimedia: highlighting opportunities for search and mining of multimedia data in social media applications. Multimedia Tools and Applications, 56, 9 - 34.

[63] Chen, Y., Liu, J., Peng, L., Wu, Y., Xu, Y., & Zhang, Z. (2024). Auto-encoding variational bayes. *Cambridge Explorations in Arts and Sciences*, *2*(1).

[64] Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

[65] Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., ... & Abbeel, P. (2016, November). Variational Lossy Autoencoder. In *International Conference on Learning Representations*.

[66] Gulrajani, I., Kumar, K., Ahmed, F., Taiga, A. A., Visin, F., Vazquez, D., & Courville, A. (2016, November). PixelVAE: A Latent Variable Model for Natural Images. In *International Conference on Learning Representations*.

[67] Van Den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016, June). Pixel recurrent neural networks. In *International conference on machine learning* (pp. 1747-1756). PMLR.

[68] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., & Metaxas, D. N. (2017). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 5907-5915).

[69] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X. (2018). Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1316-1324).

[70] Tao, M., Tang, H., Wu, F., Jing, X. Y., Bao, B. K., & Xu, C. (2022). Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16515-16525).

[71] Tao, M., Bao, B. K., Tang, H., & Xu, C. (2023). Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14214-14223).

[72] Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., ... & Guo, B. (2022). Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10696-10706).

[73] Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., ... & Chen, M. (2022, June). GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning* (pp. 16784-16804). PMLR.

[74] Bian J, Yang Y, Zhang H, Chua TS (2015) Multimedia summarization for social events in microblog stream. IEEE Trans Multimed 17(2):216–228. https://doi.org/10.1109/TMM.2014.2384912

[75] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International conference on machine learning* (pp. 8821-8831). Pmlr.

[76] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.
[77] Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., ... & Mosseri, I. (2024). Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*.
[78] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).
[79] Wang, T. C., Liu, M. Y., Tao, A., Liu, G., Kautz, J., & Catanzaro, B. (2019, December). Few-shot video-to-video synthesis. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 5013-5024).
[80] https://openai.com/index/dall-e-2/
[81] https://openai.com/index/sora/
[82] https://stablediffusionweb.com/
[83] Amato, F., Castiglione, A., Moscato, V., Picariello, A., & Sperlí, G. (2018). Multimedia summarization using social media content. Multimedia Tools and Applications, 77, 17803 - 17827.
[84] E. Sagduyu, Y., Grushin, A., & Shi, Y. (2018). Synthetic Social Media Data Generation. IEEE Transactions on Computational Social Systems, 5, 605-620.
[85] Franklin, M., Halevy, A., & Maier, D. (2005). From databases to dataspaces: a new abstraction for information management. ACM Sigmod Record, 34(4), 27-33.
[86] Curry, E., Scerri, S., & Tuikka, T. (2022). Data Spaces: Design, Deployment and Future Directions (p. 357). Springer Nature.
[87] Otto, B., ten Hompel, M., & Wrobel, S. (2022). Designing Data Spaces: The Ecosystem Approach to Competitive Advantage (p. 580). Springer Nature.
[88] Siuzdak, H. (2023). Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. ArXiv, abs/2306.00814.
[89] Suno AI. (2023). Bark. Github repository, https://github.com/suno-ai/bark
[90] Jurafsky, D., & Martin, J. H. (2024). Automatic Speech Recognition and Text-to-Speech. In Speech and Language Processing (3rd ed. draft). Retrieved May 30, 2024, from https://web.stanford.edu/~jurafsky/slp3/ .
[91] Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., & Zeghidour, N. (2022). AudioLM: A Language Modeling Approach to Audio Generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 2523-2533.
[92] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., & Tagliasacchi, M. (2021). SoundStream: An End-to-End Neural Audio Codec. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 495-507.
[93] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., & Wei, F. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. ArXiv, abs/2301.02111.
[94] Wang, S., & Székely, É. (2024). Evaluating Text-to-Speech Synthesis from a Large Discrete Token-based Speech Language Model. International Conference on Language Resources and Evaluation.
[95] Zhang, D., Zhang, X., Zhan, J., Li, S., Zhou, Y., & Qiu, X. (2024). SpeechGPT-Gen: Scaling Chain-of-Information Speech Generation. ArXiv, abs/2401.13527.
[96] Ye, Z., Ju, Z., Liu, H., Tan, X., Chen, J., Lu, Y., Sun, P., Pan, J., Bian, W., He, S., Liu, Q., Guo, Y., & Xue, W. (2024). FlashSpeech: Efficient Zero-Shot Speech Synthesis.
[97] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.
[98] https://diavgeia.gov.gr/api/help
[99] https://digital-strategy.ec.europa.eu/en/policies/strategy-data
[100] https://joinup.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe
[101] https://european-union.europa.eu/principles-countries-history/languages_en
[102] https://webgate.ec.europa.eu/etranslation/public/welcome.html
[103] https://ted-data.com/
[104] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#a-single-market-for-data
[105] https://syntheticfuture.ch/en/leverage-your-ai-with-synthetic-data/
[106] https://rumn.medium.com/yolo-data-augmentation-explained-turbocharge-your-object-detection-model-94c33278303a
[107] https://www.snowflake.com/guides/what-data-marketplace
[108] https://single-market-economy.ec.europa.eu/single-market/public-procurement/digital-procurement/public-procurement-data-space-ppds_en

[109] https://internationaldataspaces.org/
[110] https://internationaldataspaces.org/we/members/
[111] https://docs.internationaldataspaces.org/ids-knowledgebase/v/ids-ram-4
[112] https://gaia-x.eu/
[113] https://docs.gaia-x.eu/policy-rules-committee/trust-framework/22.10/
[114] https://data-spaces-business-alliance.eu/
[115] https://www.opendei.eu/
[116] https://dssc.eu/
[117] https://ishare.eu/
[118] https://deployemds.eu/
[119] https://data.europa.eu/en
[120] https://digital-strategy.ec.europa.eu/en/policies/simpl
[121] https://digital-strategy.ec.europa.eu/en/policies/data-spaces
[122] https://google-research.github.io/seanet/audiolm/examples/
[123] Bonuram S., et al. (2022). Increasing Trust for Data Spaces with Federated Learning. In: Data Spaces – Design, Deployment and Future Directions. Curry, Ed., et al (eds). Springer. Switzerland. Pp. 89-106.
[124] "Council of Europe promotes human rights through international conventions, such as the Convention on Preventing and Combating Violence against Women and Domestic Violence and the Convention on Cybercrime. It monitors member states' progress in these areas and makes recommendations through independent expert monitoring bodies.". Retrieved from: https://www.coe.int/en/web/about-us/values .
[125] Council of Europe. (1950). European Convention on Human Rights. 4 November 1950. Online version available at: https://www.echr.coe.int/documents/d/echr/convention_eng
[126] European Union. (2012). Charter of Fundamental Rights of the European Union. 26 October 2012. Publication online available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:12012P/TXT
[127] Blanke, H.-Josef., & Perlingeiro, Ricardo. (Eds.). (2018). The Right of Access to Public Information: An International Comparative Legal Survey (1st ed. 2018.). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-55554-5
[128] Floridi, Luciano. On Good and Evil, the Mistaken Idea That Technology is Ever Neutral, and the Importance of the Double-charge Thesis. Philosophy & Technology, September 2023. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4551487. Accessed on 30 May 2024.
[129] European Union. (2016). Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data (Law Enforcement Directive). 27 April 2016.
[130] European Commission. (2019). Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union. 20 May 2019.
[131] European Union. (2018). Regulation (EU) 2018/1807 of the European Parliament and of the Council on a framework for the free flow of non-personal data in the European Union. 14 November 2018.
[132] Article 1 of the Regulation (EU) 2018/1807 on the free flow of non-personal data in the EU.
[133] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R0679
[134] European Union. (2019). Directive (EU) 2019/1024 of the European Parliament and of the Council on open data and the re-use of public sector information, 20 June 2019.
[135] European Commission. (2024). A European strategy for data. Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/strategy-data
[136] European Union. (2022). Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). 3 June 2022.
[137] European Union. (2023). Regulation of the European Parliament and of the Council on Harmonised rules on fair access to and use of data (Data Act). 9 November 2023.
[138] Katulić, T., Musa, A., & Lončar, D. (2023). Understanding some of the open data challenges to data protection in the developing European legal framework. Central European Conference on Information and Intelligent Systems, 35–41.
[139] European Union. (2024). Regulation (EU) 2024/903 of the European Parliament and of the Council laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act). 13 March 2024.

[140] NCP Flanders. (2024). Interoperable Europe Act enters into force. Retrieved from: https://ncpflanders.be/news/interoperable-europe-act-enters-into-force#:~:text=The%20Interoperable%20Europe%20Act%20entered,transformation%20of%20the%20public%20sector.

[141] European Council. (2024). Interoperable Europe act: Council adopts new law for more efficient digital public services across the EU. Press release. Retrieved from: https://www.consilium.europa.eu/en/press/press-releases/2024/03/04/interoperable-europe-act-council-adopts-new-law-for-more-efficient-digital-public-services-across-the-eu/

[142] European Union. (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). 19 October 2022.

[143] European Centre for Algorithmic Transparency. (2023). FAQs: DSA data access for researchers. Retrieved from: https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en .