Version 1.0
July 2022
Status: Draft

GUIDELINES FOR ANONYMISATION OF DATA

C-DAC & STQC
Government of India
Ministry of Electronics & Information Technology
New Delhi

This page is intentionally left blank

Change History

Version No.	Release Date	Change Details
1.0	20.07.2022	Preliminary Draft

Document Metadata

S. No.	Data elements	Values
1.	Title	Guidelines for Anonymisation of Data
2.	Title Alternative	Best Practices for Anonymisation of Data for e-Governance
3.	Document Identifier	
4.	Document Version, month, year of release	Version: 1.0 (for Public Review) July 2022
5.	Present Status (Draft/ Released/ Withdrawn)	Draft
6.	Publisher	Ministry of Electronics & Information Technology (MeitY), Centre for Development of Advanced Computing (C-DAC), Pune and Standardization Testing & Quality Certification Directorate (STQC)
7.	Date of Publishing	
8.	Type of Standard Document (Standard/ Policy/ Technical/ Specification/ Best Practice /Guideline/ Framework /Procedure)	Guideline
9.	Enforcement Category (Mandatory / Recommended)	Recommended
10.	Creator (An entity primarily responsible for making the resource)	MeitY, STQC, C-DAC
11.	Contributor (An entity responsible for making contributions to the resource)	Experts from Government, R&D institutions, Industry & Academia

S. No.	Data elements	Values
12.	Brief Description	Data anonymisation is one of many privacy enhancing techniques that removes or minimises identifiability of individuals in large data sets of personal information. It reduces risk, enables entities to process, share and publish data for various purposes making it useful while optimising privacy. It is recommended that all e-governance projects apply data anonymisation as a norm and identify and record exceptions.
13.	Target Audience (Who would be referring / using the Standard)	All entities involved in processing of personal information (and subtypes) in e-governance projects. The document can also be used by private sector organisations processing personal information.
14.	Owner of approved and validated Guideline/ Framework/ Standard	Ministry of Electronics & Information Technology (MeitY)
15.	Subject (Major Area of Standardization)	Data Privacy, e-Governance
16.	Subject. Category (Sub Area within major area)	Anonymisation of Data
17.	Coverage. Spatial	India
18.	Format (PDF/A at the time of release of final Standard)	PDF
19.	Language (To be translated in other Indian languages later)	English (To be translated in other Indian languages later)
20.	Copyrights	MeitY GoI, New Delhi

Table of Contents

Terms, Definitions & Acronyms

- **Anonymised dataset** - The dataset after anonymisation technique(s) has/ have been applied on it.
- **API** - Application Program Interface
- **Attribute** - Data field, data column or variable. Name, Id and email are examples of attributes.
- **C-DAC** - Centre for Development of Advanced Computing
- **Data Archive -** A place where data is stored, but no maintenance or general usage occurs.
- **Data Archival -** Removal of the data from all active production environments and copying it to an environment where it is stored for any future need to be again in an active production environment.
- **Data destruction or purging -** The removal of every copy of a data item from an organisation
- **Data Fiduciary -** An individual, state, organization, or entity that chooses how their data should be stored, processed, and handled
- **Dataset** - A set of data records. Conceptually similar to a table in a typical relational database or spreadsheet, having rows and attributes, but could be in different formats
- **De-identification** - Process of preventing someone's identity from being revealed from data sets
- **Direct identifier** - A data attribute that identifies an individual, e.g. fingerprint, assigned to an individual. e.g. Aadhaar number.
- **eGov**- e-Governance
- **GoI-** Government of India
- **Identifiability vs Re-identifiability** - The degree to which an individual can be identified from one or more datasets containing direct and indirect identifiers, versus the degree to which an individual can be identified from the anonymised dataset(s).
- **Indirect identifier / quasi-identifier** - A data value that does not identify an individual independently but can become identifying in combination with other quasi-identifiers such as an area code, or zip code, or date of birth.
- **MeitY** - Ministry of Electronics and Information Technology
- **Metadata** - Metadata is data about data. It provides information about a particular item's content. Metadata helps us describe the technical information precisely for better understanding; it enables easier storage, access and retrieval of information and allows seamless data exchange.
- **Non-identifier** - Datasets may contain data attributes that are neither categorised as direct nor indirect identifiers. Such attributes need not undergo anonymisation.

- **NPD -** Non-Personal Data
- **PDPB Bill -** The Personal Data Protection Bill, 2019
- **Personal data** - Any information related to an identified or identifiable data subject (natural person).
- **PII -** Personally Identifiable Information
- **Processors -** Teams involved in processing the data for converting raw personal data into anonymised form.
- **Pseudonym -** An independent value through which an individual cannot be identified
- **Pseudonymisation -** The technique of replacing an identifier relating to an individual with an independent value (pseudonym) through which the individual cannot be identified.
- **Record** - Also referred to as a row. A group of information typically relating to a subject or transaction.
- **Re-identification -** Identifying a person from an anonymised dataset. Re-identification is the process by which anonymised data is linked back to its source
- **STQC** - Standardization Testing and Quality Certification
- **Third-Party** - A natural or legal person/ company that is not a data fiduciary, but processes data on behalf of another organisation typically through contractual obligations
- **UIDAI** - Unique Identification Authority of India
- **Users** - Various departments at the centre and state level, implementing agencies that will be using e-Government Standards.
- **VID -** Virtual Identifier
- **WG** - Working Group consisting of members drawn from relevant Govt. departments, subject experts from industry and Academia etc.

Disclaimer

This document is informative and advisory in nature and aims to provide guidelines to all entities involved in processing of personal information (and subtypes) in e-governance projects. The document can also be used by private sector organisations processing personal information.

Certain commercial entities, technology, or materials may be identified in this document in order to describe a concept adequately. Such identification is not intended to imply recommendation or endorsement by C-DAC, STQC and MeitY, Government of India (GoI).

While every care has been taken to ensure that the contents of this document are accurate and up to date, the readers are advised to exercise discretion and verify the precise current provisions of law and other applicable instructions from the original sources. It represents practices as on the date of issue of this document, which are subject to change without notice.

The document enlists guidelines around basic controls and is not prescriptive in nature. The readers are responsible for making their own independent assessment of the information in this document.

MeitY and/or its associated/attached offices and organizations retain the right to make changes to this document at any time, without notice. Further, MeitY and/or its associated/attached offices and organizations makes no warranty for the use of this document and assumes no responsibility for any errors which may appear in the document, nor does it make a commitment to update the information contained herein.

In no event shall C-DAC, STQC and MeitY, GoI be liable for any compensations whatsoever (including, without restriction, damages for loss of profits, business interruption, loss of information) arising out of the use of or inability to use this document.

Executive Summary

The Ministry of Electronics & Information Technology (MeitY), Government of India has entrusted Standardization Testing Quality Certification (STQC) Directorate and Centre for Development of Advanced Computing (C-DAC) Pune to formulate standards and guidelines in the areas of e-Governance. These standards/ guidelines shall ensure sharing of information and seamless interoperability of data across e-Governance applications. One of the topics/ areas under e-Governance applications is Anonymisation of Data. Accordingly, C-DAC constituted a Working Group (WG) of subject matter specialists vide letter No. CDACP/ AAI/ EGSG/ 2021/ 02/ 04 dated 18th February 2021.

The primary goal of the Working Group is to frame guidelines for anonymising data when personal information is processed and shared, especially in various e-governance applications. Terms of Reference and List of the Working Group Members are provided in Annexure 1 & 2, respectively.

The Working Group deliberated on prevailing policies, guidelines and other related literature from India and across the World. This outcome report proposes guidelines for anonymising data in e-governance and other similar projects. It includes technology options for implementation without delving into data use-case scenarios pre and post anonymisation. It also lists known data anonymity techniques and suggests practices that organisations can adopt to strengthen their privacy profile.

Structure of the Report

The Executive Summary summarises the background of setting up this WG and the purpose, scope and intended audience of these guidelines.

Chapter-1 sets the context and provides background on the exploding information landscape, processing in typical e-Governance projects and information lifecycle. It further briefly discusses the Data Anonymisation process: identifying data sources, defining datasets for anonymisation, determining what, when and how to anonymise.

Chapter-2 discusses technology options for anonymisation. It describes types of anonymisation and various anonymisation techniques, along with illustrations for each technique. It then compares the different techniques, brings out their strengths & limitations and tabulates the criteria for selection of anonymization techniques. It also covers techniques for anonymisation of specialised data, e.g. audio, video and images. A case study has also been included.

Chapter-3 identifies various stakeholders involved in the anonymisation process in an organisation. The section on Standard Operating Procedures lists down typical methodology to be followed by the organisations to anonymise datasets. A section on risk assessment is helpful in reviewing the risks continually. It provides a sample checklist to be followed for implementation and audit. It also covers the crucial aspects related to capacity building and awareness creation to fulfil the objectives of data anonymisation.

Annexure-3 gives an overview of some open-source data anonymisation tools.

Annexure-4 provides another case study where an illustrative step-by-step procedure for anonymization of a sample hospital data is explained along with screenshots.

Purpose

The purpose of these guidelines is to provide comprehensive information and practical guidance on anonymisation of data to all the intended readers and stakeholders. These guidelines shall help enhance privacy protection through case-dependent data anonymisation while processing, publishing, storing or sharing data with other entities. With the integration of multiple services, the implementation of anonymisation principles will help design privacy cognizant systems. The report provides various anonymisation techniques, supported by a few case studies and illustrations, with the aim that the stakeholders may examine and choose the techniques best suited for their applications and services. Another purpose of this report is to provide Standard Operating Procedures and checklists so that the concept of data anonymisation can be implemented conveniently by various organisations and government departments.

Another purpose of this guideline document is to supplement the data related policies/ legislations of the Government of India. While the various policies on data usage and governance would serve as the high-level documents, this document would serve as the detailed implementation-level guidelines to facilitate implementation of the data anonymisation aspects of the policies.

Scope

In line with the Working Group Terms of Reference, the document provides guidelines for “Anonymisation of Data” when processed in e-governance and associated projects. De-identification techniques that minimise risk of identifiability are commonly referred to as Anonymisation and hence the term Anonymisation gets used more generically. These guidelines are equally helpful for all organisations that seek to anonymise personal data for any reason. In addition, illustrations, use cases and practical recommendations will also help private sector organisations use data anonymisation to minimise risk.

By nature, guidelines are not mandatory but are best practices, and appropriate authorities can later decide how to evangelise them. However, to build a culture of security and privacy protection, the Working Group recommends adopting data anonymisation at scale as best practice and a norm.

Intended Readers

These guidelines would be helpful to:

- All e-governance, enterprise and institutional project owners
- Policy governance, technology architects and digital strategists
- Data protection professionals, implementers and auditors
- Personnel involved in data analysis and data-centric product development
- Technology & Service providers from India and across the World

Chapter 1: Introduction

1.1 Context

Recent developments in technology are fuelling the information revolution: our capacity to generate, gather, analyse, use and share data has increased manifold and is growing at a more than exponential rate. In addition, enhanced computing capabilities and the embedding of technology in our daily lives contribute to expanding the digital nervous system. As a result, technology and data now underpin almost all aspects of our lives.

Judicious use of technology and data helps promote innovation, empowers individuals and entities, increases the democratic quotient of our society, enhances productivity and efficiency of systems and processes, increases safety and security, and helps better understand different worlds. Also, digital transactions reveal a part of our life that gets digitised and recorded and can be used to build profiles. Such actions include the use of public services, our social media exchanges, shopping patterns, location and movement, fitness pursuits or digital content consumption behaviour.

India has seen phenomenal rise of digital adoption thus rapidly bridging digital divide in past few years: two-thirds of Indians are now Digital Nagriks, networked payments infrastructure like UPI, coordinated healthcare and vaccination platforms like CoWIN & Ayushman Bharat, identity and direct benefit transfer using Aadhaar, knowledge sharing platforms, e-Marketplace, and enhanced use of online services on a daily average. The ongoing decade has been termed “Techade”, thereby signalling the significance of tech-data in national empowerment. Tech-data systems are being envisaged as an infrastructure layer, serving as a foundation for future digital blocks. These population scale services are also being adopted by other countries.

At the same time, processing abuse poses privacy risks and other harms at an individual and collective level, violating our rights, constraining our freedoms. Unauthorised access, disproportionate surveillance, identity theft, financial frauds, blackmail & extortion using personal information, 360-degree profiling are some of the emerging challenges for society at large. In addition, large scale data breaches are rampant and present a significant challenge that requires attention from all stakeholders. The Right to Privacy focuses on empowering individuals to make informed choices in information processing and provides effective right to recourse when infringed upon.

More than 130 countries now have data protection regulations in some form to tackle the quagmire of data ecosystems: a near black-box on personal data gathering and processing by governments; law enforcement and intelligence entities; service providers; cloud and data centres; applications; websites; platforms; browsers and plugins; operating systems; network intermediaries; sensors; device manufacturers; payment processors; third parties; security solutions providers; auditors; researchers, etc. Ideally, users should be able to exercise control over their information processed across layers. Practically, empowered consumers would require informed decision making, more transparency and visibility on data processing ecosystems, curb excessive and bad faith processing to advance digital economy goals.

Parallelly, an increasing number of digital economy trade agreements (US-Japan, Japan-UK, Australia-Singapore, etc.) and multilateral arrangements like Comprehensive and Progressive Agreement for Trans-

Pacific Partnership (CPTPP), US-Mexico-Canada (USMCA), and the Regional Comprehensive Economic Partnership (RCEP), all discuss promoting cross border data flows. As a result, Data-Tech is now at par with goods and increasingly becoming the centrepiece of inter-country collaboration dialogue, especially amongst democracies. Hon'ble Prime Minister recently emphasised India's responsibility as a global "Data Power House", reiterating India's stated ambition and vision of becoming the "data processing hub" of the world.

The possibilities going forward could be both fascinating and scary. It is upon us to demarcate the desired, the acceptable and the harmful and accordingly design regulations, technology and processes to reduce risks. With the steady rise of the data-oriented world, all efforts must be made to increase data anonymisation as a positive-sum action to balance privacy protection and unhindered processing.

1.2 Regulatory Approaches

In 1980, the Organization for Economic Co-operation and Development (OECD) came up with guidelines governing the protection of privacy and transborder flows of personal data focusing on the practical implementation of privacy protection through an approach grounded in risk management. Data protection principles primarily included providing notice to data principals, limitation on collection and use of data, among others around which many global approaches of regulatory governance developed in the next three decades. Evolving from Right to be Left Alone, the umbrella of Privacy protection now includes Right to be forgotten and erased, right to access and correction, right to data portability, exercising Rights in extraterritorial processing etc. under numerous frameworks and legislations globally.

In the last decade, regulations globally have incorporated requirements to anonymise data under Data Protection laws. Regulators in the EU, UK, Singapore and other countries have released guidelines on data anonymisation, which are seeing increasing adoption in the public and private sector.

Principles of data protection do not typically apply to anonymised data sets. India's draft Personal Data Protection Bill (PDPB) 2019 explicitly mentioned that it is not applicable to the processing of "anonymised data". PDPB also provides for criminal liability for reidentifying individuals using anonymised data sets without consent. It requires data protection authority to issue "methods of de-identification and anonymisation" within a year of establishment. A case needs to be presented even for research purposes on why an anonymised data set will not be useful before processing personally identifiable information. Section 91 of the draft PDPB mentions that the Central Government may, in consultation with the Authority, direct any data fiduciary or data processor to provide any anonymised data set (obtained from personal information) or other non-personal data for better service delivery targeting or evidence-based policy formulation by the Central Government.

Kris Gopalakrishnan committee's draft on Non-Personal Data (NPD) governance framework suggests establishing rights over non-personal data, including anonymised data, coupled with community based ownership model. Thus, exercising certain rights such as recourse, access, disassociate, port, demand share becomes possible. The governance regime envisaged that all anonymised data that at the time of evaluation has not been reidentified will be governed by the NPD framework and if the data set is no longer anonymised,

it would be governed under the ambit of PDPB. NPD recommends that at the time of collecting personal data, organisations should provide a notice and option to the data principal to opt-out of processing following data anonymisation.

However, the Joint Parliamentary Committee (JPC) reviewing **Personal Data Protection Bill** in its recommendations strongly advocates to include anonymised data (and other non-personal data) under the regulatory framework of the broadened Data Protection Bill 2021. A new clause has been inserted under Section 2, that says the provisions of the Act shall apply to 2(d) **the processing of non-personal data including anonymised personal data**. The committee also recommended that a single federal regulator Data Protection Authority (DPA) shall be created that will 'regulate' both personal and non-personal data. The committee also added that both personal & non-personal data breaches need to be reported and DPA will specify on necessary steps to be taken. Section 95 allows the Central government to issue rules through notification on the steps to be taken by DPA for non-personal data breach. To ensure Transparency in processing of personal data, a data fiduciary needs to make details regarding the fairness of the algorithm or method used for processing of personal data available as per regulations. JPC further added through Section 92 that nothing in the Act shall prevent the Central government from framing any policy for digital economy, growth and misuse of data and specially provisions for handling of non-personal data including anonymised data.

Recently proposed draft of **National Data Governance Framework Policy** (NDGFP) aims to maximise data-led governance and catalyse data-based innovation through standardized data management, APIs for Whole of Government Data governance and access. It proposes setting up of India Data Management Office (IDMO) that shall prescribe rules and standards including Anonymization standards and rules for all entities (Govt and private) that will cause every Government Ministry/ Department/ Organisation to identify and classify available datasets and build a vibrant, diverse and large base of datasets for research and innovation whilst maintaining informational privacy.

Evolving the data breach reporting landscape also requires reporting incidents to regulators and government entities. Some of these proposals also include notifying breach of anonymised data sets. Given multitudes of parallel efforts, it remains to be seen how the regulatory regime shapes up to regulate non-personal data, especially anonymised data.

1.3 e-Governance Processing

Multiple e-governance projects are being implemented to improve the delivery of public services and simplify accessing them. Under the National e-Governance plan (NeGP), a concerted effort is being made to ensure 31 Mission mode projects are executed at National and State levels for the benefit of people. Projects, such as UID (Aadhaar), e-Pramaan, Banking, Insurance, Custom-Excise and Income Tax at the Central level and Education, Healthcare, Agriculture, Road Transport, e-Panchayats and Municipalities at the State level along with few integrated projects, require gathering and processing of personal information (often sensitive information) by various entities for provisioning services.

New age projects like National Health Mission, Cowin vaccination, Aarogya Setu and healthcare data, Smart cities, Payment ecosystem (Account Aggregators), upgraded Security and Surveillance apparatus and many

more such projects generate vast volumes of data. Multiple projects might need to scale to a large volume when required instantly during emergencies, law and order issues, elections, periodic collection and assessments. Digitally oriented governance models are being re-architected to bring forth the Whole-of-Government approach. India Enterprise Architecture (IndEA) was notified to help design agile frameworks for Enterprise Architecture for Ministries of GoI, the States and large public organisations. It packs several components: registries, directories, exchanges, gateways, references to engines, tools & applications, data marketplace and APIs for seamless interoperability.

Integration of services often requires cross-functional access and processing of personal information. Take an example of UMANG, a web and mobile-based application to enhance 'Ease Of-Living' with 24x7 single-point digital access to major government services from the Central Government, State/ UT Governments, local bodies and their agencies. It aims to cover Certificates, Education, Energy, Farmers, Finance and Banking, General Health, Police and Legal, Public Grievance, Ration Card, Social Justice and Empowerment, Social Security & Pensioners, Students, Tourism and Culture, Transport, Utility, Women & Children, Youth, Skills and Employment under one platform. It consists of many components coupled together for Authentication and Authorisation, Analytics, Self-care for departments, Campaign and Notification management, SMS Gateway integration, Interactive Voice Response System (IVRS), Transaction Management, Payment Gateway integration, Aadhaar integration, DigiLocker integration and CRM with multiple layers built into frontend, backend, APIs and gateways integration. Such platforms are designed keeping in mind that information flows across applications and projects owned by different departments across the country. Moreover, in some projects, data from other sources, including data gathered and processed by the private sector, must be integrated with project data for various purposes.

Thus, Government entities are the most extensive data fiduciary and accountable for lawful processing and protection of personal information. Safety and privacy protection become paramount in imparting these services to build trust in use of digital systems and platforms. Integrating large data sets containing personal information increases privacy risks during processing. Every effort must be taken to minimise risks throughout the information lifecycle. Right questions should be asked at every level of People-Process-Technology layers to identify and reduce threats and emanating risks. Therefore, all feasible measures to improve security and privacy should be followed with due care. One such Privacy-enhancing technique is "data anonymisation", which should be followed together with other technology, tools and process optimisations to better the protection profile of the organisation and reduce risks for data principals.

The use of open systems and architecture, transparent processes and agile building blocks are steps in the right direction. Ensuring privacy, safety and security, enforcement of legal and ethical principles, eliminating bias in the world driven by algorithms and codes goes hand in hand with progressive principles.

1.4 What is Data Anonymisation?

Data Anonymisation is a processing technique that removes or modifies direct and indirect personally identifiable attributes to eliminate or significantly reduce identifiability. It typically results in "anonymised data

sets" that cannot be associated with an individual. Indian draft Personal Data Protection Bill (PDPB) and the JPC draft Data Protection Act 2021 define Anonymization concerning personal data as the "irreversible process of transforming or converting personal data to a form in which a data principal cannot be identified", meeting the standards of irreversibility specified by the Authority (proposed Data Protection Authority). Data which has undergone the process of anonymization is referred to as anonymized data. The drafts also define 'de-identification' as the process by which a data fiduciary or processor may remove, or mask identifiers from personal data, or replace them with such other fictitious name or code that is unique to an individual but does not, on its own, directly identify the data principal. Typically, de-identification is one of the first steps in the process of anonymisation. Reducing the risks of identifying individuals to a sufficiently remote level is effectively deemed as anonymisation.

There are claims and counterclaims on how effective anonymisation is. The evolution of data science and technology will find more robust solutions in times to come. There are also interesting debates on if the source data set is not deleted post anonymisation, can the updated data set be considered anonymised by itself? Many use cases where gathering authentic data in easy and timely fashion is possible, might result in processing of only anonymised data sets for desired purposes, and destroy the source data so that only the anonymised version remains, and organisations perform processing on anonymised data only wherever feasible. This is possible in cases where data retention isn't critical & data can be easily gathered again if required.

Anonymisation could be done by using a single or combination of one or more privacy-enhancing techniques that help reduce the threat surface and risks emanating from the processing of large scale personally attributable data. Individuals can be re-identified from anonymised data by combining or matching multiple sources. Just removing direct identifiers does not make data sets anonymous. Linking of multiple data points can make it more likely that an individual is identifiable.

When effectively done, anonymisation can help protect the privacy rights of "data principal" balanced against reasonable purpose/ legitimate interests. It acts as a layer of defence to reduce data abuse following unsanctioned access. Providing real-like but de-identified data sets help satisfy purposes for which consent has not been collected (thus meeting the requirement of Purpose Specification Principle) and enforces the Use Limitation Principle. It also acts as a Security Safeguard since even if such data were to leak, it would not reveal the identities of any data principals by itself. As it retains data utility for certain specific purposes, it is a valuable technique for data archival. Overall, it helps reduce the organisational risk while enabling it to make the best of the available data and be accountable.

Anonymisation of personal data is possible in many circumstances, but it should not be seen as a silver bullet to tackle data privacy risks. It should also not be viewed as an auto-execute tool for compliance purposes. Data anonymisation should form part of an organisation's data governance strategy and is an important privacy-by-design approach component. We increasingly see organisations formulating Data Anonymisation policies as standalone but more typically as part of Data Protection/ Information Security policies. Practising anonymisation would be taken into account in data protection impact assessments and audits. It reduces risk

if data is breached and embodies the data minimisation principle. Overall, large scale implementation of data anonymisation would aid user confidence and trust in using digital products and systems.

The past few years have seen focused research in the field of anonymisation. The effectiveness of various anonymisation techniques is constantly challenged by increasing computing power, greater digital inclusivity and ever-improving data re-identification methods that are able to glean data from seemingly disparate systems. It seems practically impossible to conclude that a particular technique will be 100% effective in protecting the identity of data principals. New techniques and standards keep on emerging that help raise the effectiveness of anonymisation. This guidance is intended to assist with identifying and minimising the risks to data principals when anonymising data.

1.5 What, When and How to Anonymise?

Anonymisation of data can happen at multiple stages of the information lifecycle, depending on the use case. At what stage; data has to be anonymised should form part of organisational data processing strategy. The assessment of what all data processing should be minimised would determine organisational use limitations. In some instances, data sets can be anonymised at the data generation layer itself, thereby significantly reducing the risk (e.g., census, diversity data during employment, surveys, aggregated analysis). In some scenarios, data anonymisation could be inversely proportional to processing purposes (e.g., lawful processing). Generally, a higher degree of anonymisation might result in a lower use case. Determining what and what not to anonymise, at which stage of the information lifecycle should anonymisation happen, how to anonymise data should be done by the organisations in line with their objectives, referencing upgrades of regulatory regimes and emerging standards. A few use cases discussed would be helpful in determining how to approach implementation. As best practice, and in line with the principle of data minimisation, anonymisation of personal data should be done before active data processing or the earliest in the information lifecycle, wherever feasible. In future when anonymisation becomes standardised, organisations might need to reason out and record a note as to why anonymisation isn't applied.

Different industry sectors and projects would require different implementation strategies, depending on the sensitivity of data and potential harm that may be caused to data principals in the event of re-identification etc. We have rights being enforced through legal principles like the right to be forgotten, or other court rulings requiring anonymisation of data. In some cases, data will need to be anonymised when sharing with other entities for processing or publishing. Few use cases can work with anonymised data sets processing, and don't require personally identifiable data sets at all. Thus, practices need to be consolidated to develop sectoral and industry standards on adoption of data anonymisation.

1.6 Processing Purposes

Broadly organisational data processing could be categorised as:

I. Purpose based processing, which requires the organisation to clearly define all such purposes and get clear, affirmative, and explicit consent from the data principals for processing purpose.

II. Processing to fulfil a lawful disclosure request.
III. Sharing data with data-processors/third parties and other entities for processing purposes.
IV. Processing to integrate products and services with other data-tech ecosystems for the benefit of consumers.
V. Any additional processing that the organisation carries out to improve services, cross-sell, collaborate or maintain the competitive edge. In some cases where processing is experimental or short-lived, some organisations do not typically declare it as a formal purpose and collect consent against it.

The guidelines, however, do not delve into processing purposes but focus on identifying appropriate data anonymisation approaches to be in line with e-governance project's processing principles.

1.7 Typical Information Lifecycle

All the information follows a lifecycle from birth to death, from creation to deletion. Typical information lifecycle in any organisation includes:

1.7.1 Data Creation/ Gathering

The first phase of the data lifecycle for an organisation is the creation/collection of data. It can happen in multiple ways

- Data Collection: Capture of data generated by the users
- Data Acquisition: Acquiring already existing data that has been produced outside the organisation
- Data Creation: Creation of data sets either directly or in conjunction with other products and services

This data can be in many formats, e.g. text, PDF, image, Word processors, SQL database data, voice, video, and so on.

Note: During this phase, the records to be anonymised based on the legal requirements/ business requirements have to be identified. In terms of anonymisation, this phase shall be an identification phase and getting the visibility to the datasets, which requires action going forward to the next stage in the entire data life cycle.

1.7.2 Storage

After gathering data, it needs to be stored and protected by the organisation with the appropriate level of security. A robust backup and recovery process should also be implemented to ensure data retention during the life cycle.

Note: During this phase, identified datasets have to be anonymised using suitable de-identification techniques while storing the data. For instance, character masking can be enabled for storing the identity numbers such as Aadhaar, Voter ID, License etc.; generalisation techniques can be used for addresses, and so on. As per the legal requirements and business requirements, the data owner must define the parameters.

1.7.3 Usage

During the usage phase of the data lifecycle, data is used to support activities in the organisation. Data can be viewed, processed, modified, and saved. An audit trail should be maintained for all critical data to ensure that all modifications to data are fully traceable. Data may also be made available to share with others outside the organisation.

Note: During this phase, datasets to be used should be carefully investigated on a case-to-case basis by the data owner. This is also applicable for sharing data with a third party for any business usage.

For instance, while printing/ sharing the details of the data subjects for reporting/ billing/ third party screening etc., data owners should check whether any sensitive data about the data subject e.g. health records, PAN number, etc. should be anonymised before submitting the data for viewing/ processing/ modification by own usage or third party usage as per the legal requirements and business requirements.

1.7.4 Archival

Data Archival is the removal of the data from all active production environments and copying it to an environment where it is stored for any future need to be again in an active production environment.

A data archive is simply a place where data is stored, but no maintenance or general usage occurs. Then, if necessary, the data can be restored to an environment where it can be used.

Note: During this phase, data owners to ensure the following:

- Datasets are going to the backup / archival purposes as per the business/legal requirements
- Datasets retrieved from the archival need to be analysed to confirm in which form the data should be submitted for usage at the time of retrieval based on the current business/ legal requirements from time to time.

1.7.5 Destruction

The volume of archived data inevitably grows, and while one may want to save all the data forever, that is not feasible. Storage cost and compliance issues exert pressure to destroy data you no longer need. Data destruction or purging is the removal of every copy of a data item from an organisation. It is typically done from an archive storage location. The challenge of this phase of the lifecycle is to ensure that the data has been properly destroyed. Before destroying or anonymizing data, it is essential to ensure that the data items have exceeded their minimum required regulatory retention period.

Note: During this phase, data owners ensure the data destruction as per the legal and business requirements, from all the third parties/servers/ or any place where the data were earlier shared for storing/processing/modification, even if the data is anonymised.

1.8 Data Anonymisation Process

A brief overview of the Data Anonymisation Process is as follows. Detailed procedures and guidelines are provided in the subsequent chapters.

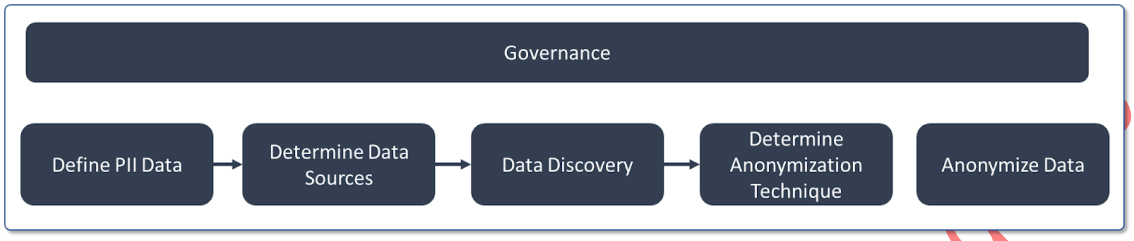


FIGURE 1: DATA ANONYMIZATION FLOW

Step 1: Identify PII and Sensitive Data

Identify Personally Identifiable Information (PII), Sensitive personal data and critical data in your department or project or software application. Example - Biometrics, Health records, financial records, authenticated services, addresses, unique identifiers etc.

Step 2: Determine Data Sources

Identify the data sources where the PII and Sensitive data is potentially stored, used or referred.

As an illustration, PII Data in the following locations needs to be anonymised.

- Application User Interface – The PII Data in the user interface should be masked.
- Database – The PII Data in the database (tables) both in RDBMS and NoSQL.
- Files – The PII data in text files (like CSV), spreadsheets (like excel), documents (like the word), pdf documents.
- Documents and Images – The files should not contain PII data in scanned documents and images (like Aadhaar, PAN).
- Log files – Application Log files containing PII Data. The log files should not print the PII Data.
- Email and SMS – PII Data in emails and SMS.
- Access log and Audit logs – The PII Data in access log and audit logs.
- Storage – The backup data in media. Network or any place should not have the PII Data.
- Prints – The PII Information should have additional security. Example Aadhaar – Currently, the Aadhaar is printed with all the details and is subjected to Data Privacy/ Security Breach.
- Multimedia – Images, Videos, Voice, Gifs etc.

Step 3: Data Discovery

- Identify the fields where the PII and sensitive data is being stored or used in the project.
- Identify all the PII fields in the application. In tables, files, logs, storage media and prints.
- The Discovery of PII Data can be manual or automated, while an automated approach is preferred over manual, to improve the accuracy of discovery.

- Define the patterns like Aadhaar, PAN.

Step 4: Determine the Anonymisation Technique

- The data Anonymisation techniques like data pseudonymisation, data redaction.
- For the purpose and the roles, determine the Anonymisation technique that would be used.
- Define Data redaction rules based on the roles. Certain information should be made available to certain roles only. For instance, the last few digits of the PAN information would be made available to high privilege users.

Step 5: Anonymize Data

- Apply Anonymisation rules for the data identified.
- Risks outstanding – Identify the data that is still not anonymised for various technical reasons. This should be reported as a part of Governance.
- For the existing applications, one-time data Anonymisation needs to be done.

Chapter 2: Techniques for Anonymisation of Data

In this chapter, various AoD techniques have been elaborated with illustrations. Their advantages, limitations and suitability for various datasets/ scenarios have been elaborated. AoD techniques for unstructured data and non-textual data including Biometric data, Multimedia data (image, audio, video, etc.), Social media data, Logs, text/ remarks fields of forms and Machine data (IoT) have been discussed. Case studies have been included to cite real-world scenarios.

2.1 Types of Anonymisation

Anonymisation is classified into two categories: static Anonymisation and dynamic Anonymisation.

2.1.1 Static or In-Place Anonymisation

It is a permanent irreversible alteration of data. The publisher anonymises the data and then publishes it for third party access. This is the one-way or release-and-forget approach of Anonymisation.

2.1.2 Dynamic Anonymisation

It is applied dynamically to the results of a query and not to the entire data set. It hides or replaces sensitive data in transit leaving the original at-rest data unaffected and offers an altered view of the real data without modifying it. Some users may only read the anonymised data whereas others may access the authentic version.

Note: Static Anonymisation is often used when the anonymisation requirements are clearly defined, and it is acceptable to do it in one go whereas dynamic anonymisation is primarily used to impose role-based privacy. It protects data in read-only scenarios and cases which does not require upfront anonymisation of all data in advance. Both types of Anonymisation can be achieved through various techniques detailed below when to use which depends on the use case.

2.2 Anonymisation Techniques

2.2.1 Attribute Suppression

The removal of an entire part of data in a dataset as per the business analysis requirements.

- Applicability

This technique can be used where an attribute is no longer required for the analysis requirement and where reidentification is not required.

- Approach
  - Delete the entire record/ column which is sensitive information/ not required for further analysis.
  - Suppression should be permanent and not just hiding/redacting the records.
  - Attribute to be selected by the data owners for ensuring the information suppressed cannot be re-identified through the data available post anonymisation.

- ## Illustration

Below table is the sample of sales information of the home appliances store. The dataset consists of customer name, customer mobile no, sales manager and amount.

Before Anonymisation

Customer Name	Customer mobile no	Sales Manager	Amount (in Rs)
Raj	98444 12345	Kumar	20000
Selvi	98777 89123	Kumar	10000
Manoj	98666 56789	Kumar	30000
Rakesh	98888 12345	Pooja	25000
Jyoti	90099 34567	Pooja	30000
Naveen	98456 12345	Pooja	30100

Table 2.a: Before Anonymisation

For analysing the sales efficiency of the respective sales manager, customer details are irrelevant and need to be protected from exposing the details to third parties. Here, the Anonymisation technique of attribute suppression can be applied here by completely removing the columns of customer name and customer mobile no.

Post Anonymisation:

Sales Manager	Purchase Amount (in Rs)
Kumar	20000
Kumar	10000
Kumar	30000
Pooja	25000
Pooja	30000
Pooja	30100

Table 2.b: Post Anonymisation

- ## Notes
  - There is a possibility of creating a “derived attribute” that provides the information in an indirect way. For instance, “Time spent in a premise” can be derived from the “Entry time” and “Exit time”. While

suppressing the attribute, it is also necessary to look for the possibility of removing derived attributes, if required.

- This is one of the easiest and the strongest type of anonymisation techniques because the possibility of reidentification is almost zero in this technique.

2.2.2 Character Masking

The change of the characters of a data value to a constant symbol or character (e.g., "x"). Masking is typically done for the partial data value.

- Applicability

This technique can be used where the data is a string of characters and hiding some part of it is sufficient to provide required anonymity.

- Approach
  - Based on the type and nature of chosen data value, replace the suitable characters with a constant symbol.
  - On a case-to-case basis, the number of characters considered for masking may vary accordingly. For instance, a fixed number of characters may be replaced for account number/ card number, or a variable number of characters for username/ email address.

- Illustration

Aadhaar number masking is one of the live examples of character masking. The below table contains the information of the name and the Aadhaar card number. As per the UIDAI guidelines, the Aadhaar card number to be masked for privacy requirements. The dataset consists of names and Aadhaar card numbers.

Before Anonymisation:

Name	**Aadhaar card number**
Rakesh	2345 6789 1234
Raj	5678 1234 5678
Celin	2345 1234 9123

Table 2.c: Before Character Masking Anonymisation

Here, the Anonymisation technique of character masking can be applied.

Post Anonymisation:

Name	**Aadhaar card number**
Rakesh	xxxxxxxx 1234
Raj	xxxxxxxx 5678

Name	Aadhaar card number
Celin	xxxxxxxx 9123

Table 2.d: Post Character Masking Anonymisation

- Notes
  - During the anonymisation, the processor needs to ensure the length of anonymised data provides information about the original data.
  - Subject matter expertise is required for partial masking to ensure the right characters are masked.
  - Sometimes, the length of data may also be increased to meet up the standard-length requirement of the anonymised data.
  - This is a unique technique in which masking allows data subjects (owners of data) to recognize their own data.

2.2.3 Pseudonymisation /coding

- Description

  It is a technique of replacing information (which is an identifier to relate with an individual) with a pseudonym (i.e., an independent value through which an individual cannot be identified). The original data values are maintained securely and can be retrieved for mapping back to the pseudonym, whenever it is required. It is also referred to as 'coding'.

- Applicability

  This technique can be used, where no information of the original data value shall be shown.

- Approach
  - Different pseudonyms may be used to represent the same individual in different datasets to prevent mapping of the different datasets.
  - Replace the respective data values with random made-up values.
  - The made-up values should be unique and should have no relationship to the original data values (no direct/ indirect identifier to the original value).

- Illustration

  Aadhaar (Virtual ID) VID is one of the classic examples of this technique, which is implemented countrywide. VID is a temporary, revocable 16-digit random number mapped with the Aadhaar number. VID can be used in lieu of the Aadhaar number whenever authentication or e-KYC services are performed. Authentication may be performed using VID in a manner similar to using an Aadhaar number. It is not possible to derive an Aadhaar number from VID. The dataset consists of name, Aadhaar number, Address and Mobile number.

Before Anonymisation:

Name	Aadhaar Number	Address	Mobile number
Anwar	2345 6789 1234	308, GHT apartment, Sector 1, Mumbai	98000 00000
Bala	4567 8912 3456	412, SR road, Sector 12, New Delhi.	98000 11111
Ram	3456 7890 5678	44, TK road, Alwarpet, Chennai.	98000 22222

Table 2.e: Before Pseudonymisation

Identity database is created as seen below. Here, VID is the pseudonym value mapped with the original data value i.e., Aadhaar number, which needs to be secured.

Virtual ID (VID)	Aadhaar Number
1111 2222 3333	2345 6789 1234
4444 5555 6666	4567 8912 3456
7777 1111 2222	3456 7890 5678

Table 2.f: Mapping of pseudonym (VID) with the original data value (Aadhaar number)

Further, VID is mapped with the other data sets associated with the original data value as seen below:

Post Anonymisation:

Name	Virtual ID (VID)	Address	Mobile number
Anwar	1111 2222 3333	308, GHT apartment, Sector 1, Mumbai	98000 00000
Bala	4444 5555 6666	412, SR road, Sector 12, New Delhi.	98000 11111
Ram	7777 1111 2222	44, TK road, Alwarpet, Chennai.	98000 22222

Table 2.g: Post Pseudonymisation

- Notes
  - When allocating pseudonyms, ensure not to re-use pseudonyms that have already been utilised (especially when they are randomly generated).
  - Security controls (including administrative and technical ones) should also be used to protect the identity database for mapping back to the original value.
  - For added security regarding the identity database, double coding can be used.

2.2.4 Data Swapping

- **Description**

  Data swapping or shuffling is the technique where attributes in a dataset are rearranged in such a manner that they do not match with the original records. fields in a record are swapped to fields of another record. Data swapping is an irreversible technique where fetching original data is nearly impossible. Swapping should be done on recognizable attributes, such as date of birth, it can make a huge impact on Anonymisation.

- **Applicability**

  Data-swapping is a best-suited data transformation where the underlying statistics of the data is to be preserved even after the Anonymisation of data and when analysis of relationships between attributes at the record level is not needed

  The result guarantees the privatization of the original dataset while providing data for accurate statistics.

- **Generic Approach**
  - Identify which attributes to swap
  - Select any two random records and swap them for selected attributes
  - Select any swap algorithms and apply it till you receive a true swap.

- **Illustration**

  The sample dataset contains Web Search logs with attributes UserID: stores the ID of the user. 'QueryTEXT' stores the actual query executed by the user. 'QueryTime' the date and time of the query execution, 'CURL' The actual URL that the user clicked on after querying. In the experiment, the swap attributes include the QueryText and URL hence who accessed what is privatized but statistics like how many times the particular site is accessed or how many times user 7 has accessed the web remains accurate.

  Before Anonymisation:

UserID	QueryText	QueryTime	URL
1	shopping website	06/13/2017 16:08	www.shop.com
2	PM of India	01/14/2016 11:18	www.pmindia.com
3	Olympics 2020	07/12/2014 22:38	www.olympics.com
4	Covid Vaccination	08/11/2012 12:58	www.vaccination.com
5	Mutual Funds	01/14/2020 22:10	www.mutualfund.com
6	Kids Toys	01/14/2013 20:08	www.toys.com
7	CBSE Results	01/14/2016 12:08	www.results.com

UserID	QueryText	QueryTime	URL
8	Data Privacy	01/14/2016 11:08	www.privacy.com

Table 2.h: Before Data Swapping

Here we will swap the URLs and QueryText so that who accessed what will be privatized but the statistics of data remains accurate.

Dataset Post Anonymisation: values after swapping

UserID	QueryText	QueryTime	URL
1	shopping website	06/13/2017 16:08	www.mutualfund.com
2	Olympics 2020	01/14/2016 11:18	www.shop.com
3	PM of India	07/12/2014 22:38	www.olympics.com
4	Data Privacy	08/11/2012 12:58	www.vaccination.com
5	Covid Vaccination	01/14/2020 22:10	www.privacy.com
6	Kids Toys	01/14/2013 20:08	www.results.com
7	CBSE Results	01/14/2016 12:08	www.pmindia.com
8	Mutual Funds	01/14/2016 11:08	www.toys.com

Table 2.i: Post Data Swapping

- Notes
  - Many times, only removing PII is not sufficient when privatizing a dataset; other anonymization techniques have to be considered. In such cases, data swapping can be used to make it more complex for re-identification and co-mingling attacks.
  - Data swapping can be done on selected attributes while keeping other attributes unaffected.
  - Sometimes swapping creates unusual impractical combinations of values, e.g. a female disease is exchanged with that of a male.
  - High computation resources are required for fully swapping large datasets.

2.2.5 Record Suppression

- Description

Record suppression means the removal of an entire record in a dataset. This technique affects multiple attributes at the same time, unlike other techniques which work on a single attribute.

- ## Applicability

  This technique can be applied when some of the existing data records in the dataset do not serve the purpose of the data evaluator in any way but do contain identifying information or which are unique in the dataset or which do not meet criteria such as k-anonymity.

- ## Generic Approach

  - Delete the entire record which has sensitive information and does not meet criteria such as k-anonymity.
  - Suppression should be permanent and not just hiding/redacting the records.

- ## Illustration

  Below table is the sample dataset which contains a person name, age, address and city. Consider that a person's name is already pseudonymized and address is generalized to the area name.

  ### Dataset before Anonymisation:

Sr no	Name	Age	Address	City
1	Ramesh	25	Andheri	Mumbai
2	Satish	26	Juhu	Mumbai
3	Manisha	35	Dadar	Mumbai
4	Mahesh	35	Mahalakshmi	Mumbai

Table 2.j: Before Record Suppression

But suppose only one residential unit is present in the Dadar area, in that case, the exact address can be found. In such cases where a record becomes identifiable in some way even after Anonymisation, record suppression can be used to remove such records. Another example could be certain foreign citizens' information may have to be excluded from the data set for regulatory requirements.

Dataset Post Anonymisation:

Sr no	Name	Age	Address	City
1	Ramesh	25	Andheri	Mumbai
2	Satish	26	Juhu	Mumbai
4	Mahesh	35	Mahalakshmi	Mumbai

Table 2.k: Post Record Suppression

- ## Notes

  - It can be applied before or after any other Anonymisation techniques have been applied (e.g. generalisation)

- Record suppression can affect the data analysis in terms of statistics such as average, sum, median etc.

2.2.6 Generalisation

- **Description:**
  Generalization involves transforming values of quasi-identifiers into more general values to make personal records unidentifiable from several people (records). It is a process of replacing the individual values of attributes with a broader range or category (less precise value) but semantically consistent value. For example: converting a person's age into an age range, or a precise location into a less precise location.

- **Applicability:**
  For values that can be generalised and still be useful for the intended purpose.

- **Approach:**
  - The identifiers must be removed. The quasi-identifiers are identified and then generalized to anonymise the data.
    - Numerical attribute values are transformed into range values. Fig. 2.a shows the Taxonomy Tree for Age attribute. In practice, instead of inserting range, attribute values are replaced by random values. For example, the attribute 'age' with a value of 22 in a dataset can be generalized to a random value in the range [20-25].
    - categorical attribute values are transformed into superordinate values. Taxonomy trees are usually employed to describe hierarchies of categorical attributes. Fig. 2.b shows the taxonomy tree of Marital Status.
  - The sensitive information should not be removed or modified because it could be a critical attribute for analysis.
  - The two main types of data generalization are automated generalization and declarative generalization.
    - Declarative generalization involves manually deciding how large your data bin sizes will be in any given scenario.
    - Automated generalization uses algorithms to determine the minimum amount of generalization or distortion required to ensure proper privacy while retaining accuracy. More details of this approach would be covered in k-Anonymisation.

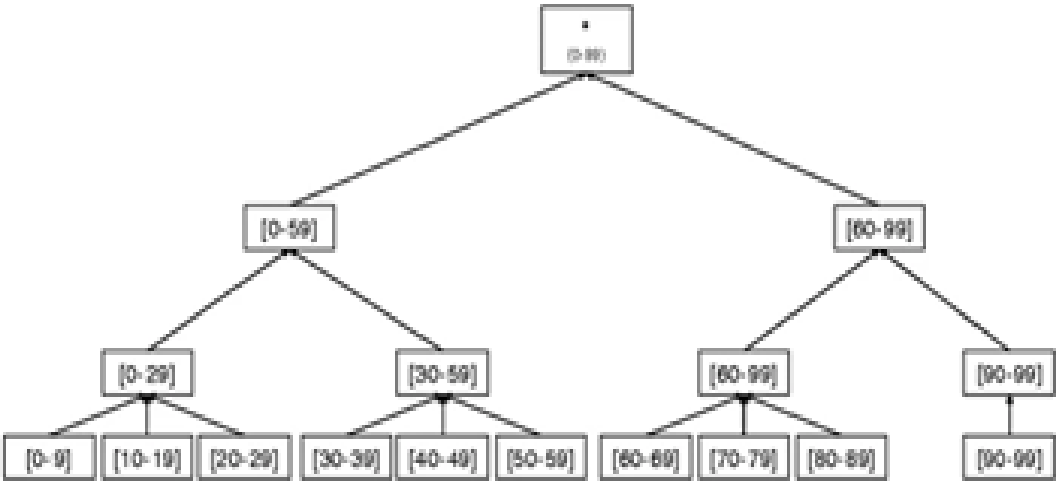


Fig. 2.a Taxonomy Tree of Age attribute

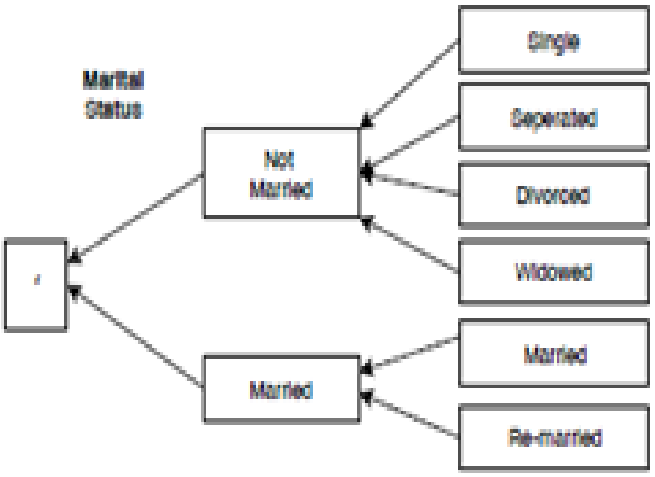


Fig. 2.b Taxonomy Tree of Marital Status

- Illustration

In this example, this dataset contains a person name (which has already been pseudonymised), Aadhaar number is masked, age in years, residential address and sensitive data (Medical report). The age is generalized in the bucket of 10 years ([20-30] and [30-40] years). For the address, one possible approach is by removing the locality details and retaining only the city name. As we have applied a single strategy for records across each attribute, it is a global recording strategy. With this generalization, we have 3 records from Pune and 3 records from Hyderabad. In most practical scenarios, the age can be generalized to a random value in the range [20-30], to be semantically consistent with the data type.

Before Anonymisation (PII's removed or masked)

Name	Aadhaar No.	Age	Address	Medical Report
Raj	*	24	Sivam Society, Shivajinagar, Pune	Negative
Selvi	*	31	MindSpace Madhapur, Hyderabad	Positive
Manoj	*	34	M.G. Road, Secunderabad, Hyderabad	Positive
Rakesh	*	29	St. Patrik Society, Hadapsar, Pune	Negative
Jyoti	*	23	Tranquility, Hadapsar, Pune	Negative
Naveen	*	37	Afzalgunj, Hyderabad	Negative

Table 2.l: Before Generalisation

Post Anonymisation using Generalization (Global recoding)

Name	Aadhaar No.	Age	Address	Medical Report
Raj	*	[20-30]	Pune	Negative
Selvi	*	[30-40]	Hyderabad	Positive
Manoj	*	[30-40]	Hyderabad	Positive
Rakesh	*	[20-30]	Pune	Negative
Jyoti	*	[20-30]	Pune	Negative
Naveen	*	[30-40]	Hyderabad	Negative

Table 2.m: Post Generalisation

- Notes
  - Design the data ranges with appropriate sizes. Data ranges that are too large may mean that the data may be "modified" very much. For example: Age range could be (with 5 years difference: [20-25], [25-30], …), with 10 years difference [20-30]). Ranges that are too small may mean that the data is hardly modified and therefore still easy to re-identify. For example: Age range with 2 years ([20-22], [22-24], …,), etc.
  - Generalization inevitably causes information loss. Higher generalization increases the difficulty for the attacker to infer sensitive data or re-identification of individuals. This technique is also referred to as recoding. There are two types of recoding: Local and Global recoding. In global recoding, the same level of generalization is applied to all values of an attribute in the table. In Local recoding, different

levels of generalization are applied to different values of an attribute in the table. The selection of the level of generation is challenging and it depends on the domain, data utilization and publicly available datasets.

- If k-anonymity is used, the k value chosen will affect the data ranges too. Also, depending on the data utility, automatic selection of generalization level would be needed. For certain similar groups of records, different local recoding strategies need to be applied, so that information loss is reduced.
- As each attribute is generalized separately, it loses the correlations between the different data table attributes. This would result in an obstacle to the efficient analysis of the attribute correlations. Hence, this method is not suitable for critical analytics applications.

2.2.7 Data Perturbation

- **Description**

  Data perturbation is a technique that adds 'noise' to the original databases to protect individual record confidentiality. Mostly adapted for continuous numerical attributes. It attempts to preserve the privacy of the data by modifying (replacing) values of the sensitive attributes using a randomized process. The goal of data perturbation algorithms is to optimize the data transformation process by maximizing both data privacy and data utility. Data perturbation includes a wide variety of techniques including (but not limited to): additive, multiplicative, matrix multiplicative, etc.

- **Applicability**

  For quasi-identifiers (typically numbers and dates) which may potentially be identified when combined with other data sources and slight changes in value are acceptable. This technique should not be used where data accuracy is crucial. As the original (private) dataset is perturbed and the result is released for data analysis. The intent of data perturbation techniques is to allow legitimate users the ability to access important aggregate statistics (such as mean, correlations, etc.) from the entire database while 'protecting' the individual identity of a record. For instance, in a simplified case of sales data, a legitimate system user may not be able to access what a particular individual purchased from a store on a given day, but that same user could determine the total sales volume for the store on the same day.

- **Generic Approach**

  It depends on the exact data perturbation technique used. These include rounding and adding random noise. The example in this section shows base-x rounding.

  The process is explained as follows:

  I. Identify the confidential and non-confidential attributes for the original database, say **U**.
  II. The data perturbation process will create a perturbed database **P** (based on **U**) that has all instances with all the attributes.
  III. The non-confidential attribute values for all instances in database **P** will have the same value as in database **U**.
  IV. All the confidential attribute values for all instances in database **P** will be perturbed (modified) from the value in the corresponding instance in database **U.**

V. A random value is added to an attribute value of a record in the original database. These random values are drawn from a certain distribution (uniform, Gaussian, etc).
   a. The perturbation process can be more evolved using the statistical relationships of database **U.** These relationships include the mean values for confidential attributes, the measures of covariance between confidential and non-confidential attribute sets, and the canonical correlation between these attribute sets.
   b. Given these statistical properties of **U**, a multivariate normal distribution function can be constructed for each instance. Then, a multivariate random number generator generates the new attribute values for each entry in the perturbed database P.
   c. This is repeated for all instances.

- ### Illustration

In this example, the dataset contains information to be used for possible linkage between a person's age, BMI - Body Mass Index and severity in CoV-2 infection. It is assumed that user identifiers are already removed or pseudonymized. The rounding is applied to Attributes "Age, Height and Weight". The sensitive element "Illness" is kept as-is. The rounding is done using Base-5 for Height and Weight, while Base-3 is used for Age.

Patient	Age	Height (cm)	Weight (kg)	Severity of CoV-2 illness
1	24	160	64	Asymptomatic
2	29	177	79	Asymptomatic
3	48	158	61	Mild Illness
4	51	173	76	Moderate
5	87	169	92	Critical illness
6	35	152	66	Asymptomatic

Table 2.n: Original dataset before data perturbation

Anonymised Dataset (using Perturbation)

Patient	Age	Height (cm)	Weight (kg)	Severity of CoV-2 illness
1	24	160	65	Asymptomatic
2	30	175	80	Asymptomatic
3	48	160	60	Mild Illness

Patient	Age	Height (cm)	Weight (kg)	Severity of CoV-2 illness
4	51	175	75	Moderate
5	87	170	90	Critical illness
6	36	150	65	Asymptomatic

Table 2.o: Dataset after data perturbation

Note: for base-x rounding, the attribute values are either rounded-up or rounded-down to the nearest multiple of x using an unbiased random rounding function. For each cell, generate a random uniform number u between 0 and 1, if (u < res(x)/b) entry is rounded up, otherwise rounded down, where res(x) is the residual of round base b.

- Note

  There are multiple ways in which perturbation could be done. The method evolved over the period. Later in this guideline report, differential privacy would be discussed.

2.2.8 Synthetic Data

- Description

  This technique is slightly different as compared to the other techniques described in this Guide. Synthetic data is described as artificially (algorithmically / programmatically) generated data that approximates original data. It retains the underlying structure and statistical distribution of the original data. Synthetic data does not rely on masking or omitting the original data. The main idea behind synthetic data generation is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data, which is then released instead of the original (real) data. Hence, the risk of re-identification is reduced.

- Applicability

  Typically, when a large amount of data is required for system testing, the actual data cannot be used and yet the data should be "realistic" in certain aspects, like format, the relationship among attributes, etc. Synthetic data can be used as a drop-in replacement for any type of behaviour, predictive, or transactional analysis. Synthetic datasets can be more openly published, shared, analysed, without revealing actual individual information.

- Generic Approach

  Given original dataset containing n records. Each row is an individual's record containing k numerical or categorical attributes (such as demographics, dates, sensitive attributes). The synthetic data generators take this original dataset as input and construct a statistical model. The generators identify the density function of attributes in the original data and estimate the parameters of these density functions. The model captures the statistical properties of the complete data (population). This model is used to generate

synthetic datasets that are statistically similar to the original dataset. The synthetic data can be generated in any volume, as needed and can be shared.

Organizations can choose to produce partially or fully or hybrid synthetic datasets. In fully synthetic data, the data is completely synthetic and doesn't contain original data. In partially synthetic data, synthetic data replaces only values of the selected sensitive attribute with synthetic values. The original values are replaced only if it possesses a high risk of disclosure. Disclosure risk is higher in partially synthetic data compared to fully synthetic data as it contains original data along with imputed synthetic data. The hybrid synthetic data is generated using both original and synthetic data. For each record of original data, the nearest record in the synthetic data is chosen and both are combined to form hybrid data. For different data types (text, media (video, image, sound), and tabular), synthetic data can be generated.

- ### Illustration

The sales department of a Supermarket keeps a daily track of sales of each product category, along with some of the customer details. It aims to promote particular products and provide discounts on various categories to different age groups. They want to provide this data to an external data analyst firm. As the data contains PII such as customer mobile number and sensitive information such as purchasing details. The real data cannot be shared. The third-party firm needs to create synthetic data and simulate various discount and pricing models.

The illustration displays data of users and the product category, purchased by them. Table 2.q - displays the count of customers in each age group. We can learn the joint probability distribution in an original dataset to generate new synthetic data, as shown in Table 2.r. A detailed discussion of statistical measures is beyond the scope of this Guide.

User	Age	Product Category
1	25	Home Appliances
2	35	Grocery
3	25	Apparels
4	30	Apparels
5	30	Home Appliances
6	35	Home Appliances
7	20	Apparels
8	33	Apparels
9	40	Grocery

User	Age	Product Category
10	27	Apparels

etc.

Table 2.p: Original Dataset before anonymisation

Product Category	Age Group	Customer Count
Home Appliances	[20 - 30]	3
Home Appliances	[30 - 40]	23
Home Appliances	[40 - 50]	14
Home Appliances	[50 - 60]	2
etc.		
Grocery	[20 - 30]	9
Grocery	[30 - 40]	128
Grocery	[40 - 50]	27
Grocery	[50 - 60]	19
etc.		
Apparels	[20 - 30]	123
Apparels	[30 - 40]	75
Apparels	[40 - 50]	30
Apparels	[50 - 60]	10

Table 2.q: Stats obtained from the original data set

Name	Age	Product Category
1	21	Apparels

Name	Age	Product Category
2	20	Apparels
3	50	Grocery
4	36	Grocery
5	45	Home Appliances
6	41	Home Appliances
7	30	Grocery
8	38	Apparels
9	32	Apparels
10	26	Apparels

etc.

Table 2.r: Synthetic dataset for one-day post anonymisation

- Notes
  - Since the released data is completely artificially generated and doesn't contain original data, the risk of data leakage is reduced but eventually, the truthfulness of the data is lost.
  - Depending on the scope of application and the administrative controls, fully or partially or hybrid synthetic data can be generated.
  - When applying this technique, outliers may need additional attention. For testing purposes, outliers are often very valuable, but outliers in the synthetic data may also indicate certain outliers within the original dataset. It is therefore recommended to create outliers in synthetic data intentionally and independent of the original data.
  - A synthetic data generator can be constructed to represent and generate a wide variety of simple-to-complex structured data sets including not only traditional tabular but also grammar and graph-based data sets. Data generation constraints can be captured in a well-defined description language. The synthetic data generator can be designed to execute efficiently and run in parallel to generate very large data sets.
  - Synthetic data has limitations as an approach because perfectly preserving both privacy and data value in a single dataset is mathematically challenging. A better solution here is to use differential privacy in combination with synthetic.
  - Differentially private synthetic data takes the chance of identification to a much safer level than vanilla synthetic data. Differentially private algorithms or DP mechanisms are randomized algorithms that add noise at key points. There are recent advances in generating synthetic data using the conditional

generative adversarial networks (GAN) framework. A detailed discussion of the GAN framework is beyond the scope of this Guide.

2.2.9 Data Aggregation

- **Description**

  Data is gathered to be combined into a comprehensive summary for the data analysis process. For example, raw data can be aggregated over a given time period to provide statistics such as average, minimum, maximum, sum, and count. After the data is aggregated and written to a view or report, you can analyse the aggregated data to gain insights about particular resources or resource groups. Typically, you'll be aggregating data in order to process that data together. Data aggregation by hand can be time-consuming and draining, so most analysts rely on data aggregation tools. Data aggregation tools are used to combine data from multiple sources into one place, in order to derive new insights and discover new relationships and patterns—ideally without losing track of the source data and its lineage.

- **Applicability**

  Data aggregation tools allow you to look beyond the two-dimensionality of a row and column tool like Excel. For example, you can apply calculations across categories, and then use the resulting high-level summary information to present overall statistics. You might want to use data aggregation tools to bring together data from your sales regions, product categories and customer trouble tickets, all by time. It's time for aggregating data when you need to look beyond a typical data layout and begin using the data.

  Data Aggregation Examples by Industry

  Most industries work heavily with data, so most industries also use data aggregation in some form. Here are how these tools are being used within several industries.

  i. **Financial industry**

  Financial analysts are constantly focusing on staying up-to-date on industry and company trends.

  ii. **Healthcare industry**

  It's crucial to monitor healthcare developments and trends to create innovative developments and correctly diagnose patients with the best information. They can use data aggregation to help maintain transparency and trust between the healthcare industry and patients.

  iii. **Marketing industry**

  Marketing professionals need these aggregation tools to gather market trends to be able to determine current needs for their businesses. Businesses also need to see how campaigns are performing, and these tools help the marketing teams determine how their campaigns are working with the customers.

  iv. **Generic Approach**

  A detailed discussion of statistical measures is beyond the scope of this Guide, however typical ways include using totals or averages, etc. It might also be also useful to discuss with the data recipient about the expected utility and find a suitable compromise.

- Illustration – Data Aggregation

The below table is the sample of sales information of the home appliances store. The dataset consists of brand name, purchase value and store name.

Brand name	**Purchase value (in Rs)**	**Store name**
Brand A	20000	Store A
Brand B	10000	Store A
Brand A	15000	Store A
Brand B	25000	Store A
Brand A	30000	Store A
Brand B	9000	Store B
Brand A	7000	Store B
Brand B	19000	Store B
Brand A	12000	Store B

Table 2.s: Original Dataset before anonymisation

For performing analysis by an external consultant on the purchase value of different ranges for making strategic decisions, aggregated data is sufficient as seen in the table below:

Purchase value (In Rs)	**No. of products sold**	**Total value of the products sold in the price range (in Rs)**
0-10000	3	26000
10000-20000	4	66000
20000-30000	2	55000

Table 2.t: Dataset post data aggregation

2.3 Measure of data privacy

In this section, we shall see a sample for measuring the level of data privacy done on the data sets. This will help the organizations to measure the level of anonymisation that has been done to ensure the privacy of the data.

2.3.1 k-Anonymity

- **Description**

  K-anonymity (and similar extensions to it like L-diversity and T-closeness) is a measure used to ensure that the risk of threshold has not been surpassed, as part of the anonymisation methodology. K-anonymity is one of the Anonymisation approaches proposed by Samarati and Sweeney [2002][13] that each record in a dataset cannot be distinguished with at least another (k-1) record. This is achieved by the projection of the quasi-identifiers of the dataset after a series of anonymity operations (e.g. replace the specific value with general value). K-anonymity assures that the probability of uniquely representing an individual in the released dataset will not be greater than 1/k.

- **Application**

  To confirm that the anonymisation measures put in place achieve the desired threshold against linking attacks. Mostly suitable for data publishing.

- **Approach**

  There are two ways to achieve k-anonymity, namely global recoding and local recoding. In global recoding, k-Anonymisation is achieved by generalizing all the records to the same level of generalization for an attribute. As all records are generalized to the same level, the global recording approach is time-efficient but the loss in data utility is high. To overcome this loss, local recoding is preferred. Indeed, in local recoding, k-Anonymisation is achieved by generalizing different records to different levels. Different approaches are proposed in the literature to balance the trade-off between utility and privacy.

- **Illustration**

  For example, the set <Age, Gender, Pincode> makes a QID set, which needs to be generalized in such a way the k-anonymity is achieved. Using the Global recoding strategy, the k=3 anonymity is achieved in the anonymised table, which can be published.

Name	Age	Gender	Pincode	Medical Report
Raj	24	Male	411005	Negative
Selvi	31	Female	500081	Positive
Manoj	34	Male	500003	Positive
Rakesh	29	Female	411028	Negative
Jyoti	23	Female	411040	Negative
Naveen	37	Male	500012	Negative

Table 2.u: Dataset before anonymisation

Name	Age	Gender	Pincode	Medical Report
Raj	[20-30]	*	411***	Negative
Selvi	[30-40]	*	500***	Positive
Manoj	[30-40]	*	500***	Positive
Rakesh	[20-30]	*	4110**	Negative
Jyoti	[20-30]	*	4110**	Negative
Naveen	[30-40]	*	500***	Negative

Table 2.v: Anonymised dataset (k-anonymised with k =3)

- Notes
  - Achieving optimal k-anonymity is an NP-hard problem for multidimensional data. Even k=2 is difficult to achieve with multidimensional data.
  - The amount of generalization and selection of QID sets depends on the amount of publicly available data. Due to uncertainty in data that is going to be published, the selection of the number of quasi-identifiers can also vary, the approach should take into consideration of such quasi-identifiers sequence in advance [12].
  - Determining the value of k itself could be challenging. The Health Insurance Portability and Accountability Act (HIPAA) in the US has recommended k=20,000 as a standard value.

2.4 Comparative Analysis on AoD techniques

Summary and Recommendation of the Techniques. the attribute type: PII, QID: Quasi-identifiers, S: Sensitive, NC: Non-confidential

Technique Name	Applicability	Attribute type	Data Type
Attribute Suppression	Attribute is not required in the anonymised dataset	All	Any
Record Suppression	Presence of outlier records	N.A. (applies across entire record, hence all attributes affected)	Any
Character Masking	Masking some characters in an attribute provides sufficient anonymity	Direct identifier	String

Technique Name	Applicability	Attribute type	Data Type
Pseudonymisation	Records still need to be distinguished from each other in the anonymised dataset but no part of the original attribute value can be retained	Direct identifier	String, Numbers
Generalisation	Data publishing	PIIs are removed; QID are generalized; S and NC as-is;	Tabular data
Swapping	No need for analysis of relationships between attributes at the record level	All	Any
Data perturbation	Slight modification to the attributes is acceptable	PIIs are removed; Add perturbation to QIDs; S and NC as-is;	Numerical
Synthetic data	Artificially generated data that approximates original data	All	All
Data aggregation	Individual records are not required and aggregated data is sufficient	Indirect identifiers	Numbers, String, Relative data

Table 2.w: Comparative Analysis of data anonymisation techniques

2.5 Strengths and Limitations of AoD Techniques

Technique Name	Strengths	Limitations
Attribute Suppression	Simple and strongest Anonymisation technique	Re-identification possibility is very low (in case of any business requirements).
Record Suppression	Easy to implement and one of the strongest Anonymisation technique	Suppression of a record can impact the data analysis in terms of statistics
Character Masking	Masking allows data subjects (owners of data) to recognize their own data.	Subject Matter expertise required to do the masking of specific characters. Re-identification is quite easy.
Pseudonymisation	Best technique where re-identification required for one-to-one mapping.	Secure management of the original raw data required.

Technique Name	Strengths	Limitations
Generalisation	Anonymisation of real data by broader categorization and range. Truthfulness of the data is preserved. QIDs are generalized. With local recoding, high utility can be achieved.	Higher generalization impacts the utility, while preserving privacy. Identifying the level of generalization is challenging. Risk from linkage attacks persists, as it is uncertain data that would be available in public.
Swapping	Can be done on some of the attributes keeping the others intact.	In random swapping, there is a probability of getting the same value as the original value. High computing resources required for large datasets
Data perturbation	Does not require knowledge of the distribution of other records in dataset	This technique cannot be used where data accuracy is crucial. A small base may lead to weak Anonymisation
Synthetic data	Large amount of data can be synthetically created. Useful for system testing and can be openly published, shared, analysed.	Risk of re-identification is significantly reduced but eventually the truthfulness of the data is lost.
Data aggregation	Provides a comprehensive summary of data. As it is at aggregate level, risk of re-identification does not exist.	Utility might be hampered and might not be useful for applications where data is needed for analysis.
k-anonymity	A measure to ensure the risk of threshold has not been surpassed, as part of the anonymisation methodology. Most suitable for data publication.	Does not provide the guarantee of privacy. Risk of linkage attack persists.

Table 2.x: Strengths and Limitations of anonymisation techniques

2.6 AoD techniques for Specialized Data (Speech, Video, Images, etc.)

In the current environment, we are confronting many big data breaches that necessitate governments, organisations, and companies to reconsider privacy. In contrast to that, almost all breakthroughs in Machine Learning come from learning techniques that require a large amount of training data. Besides, research institutions often use and share data containing sensitive or confidential information about individuals. Improper disclosure of such data can have adverse consequences for a data subject's private information, or even lead to civil liability or bodily harm.

The development of formal privacy models like Differential Privacy was helping in solving the problem. Thus, there is an increasing number of organisations and companies that are applying Differential Privacy to protect

sensitive information, such as personal information, user's events, individual's real-time location, as mentioned in this post: A High-level Introduction to Differential Privacy. There is even an open-source differential privacy project for executing differential privacy queries on any standard SQL database.

In short, Differential Privacy permits:

- Companies access a large amount of sensitive data for research and business without privacy breaches.
- Research institutions can develop differential privacy technology to automate privacy processes within cloud-sharing communities across countries. Thus, they could protect the privacy of users and resolve data sharing problems.

Differential privacy (DP) is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. According to this mathematical definition, DP is a criterion of privacy protection, which many tools for analysing sensitive personal information have been devised to satisfy.

2.6.1 What does it guarantee?

- Differential privacy mathematically guarantees that anyone seeing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis. [15]
- DP provides a mathematically provable guarantee of privacy protection against a wide range of privacy attacks (include differencing attacks, linkage attacks, and reconstruction attacks).

2.6.2 What does it not guarantee?

DP does not guarantee that what one believes to be one's secrets will remain secret. It's important to identify which is general information and which is private information to get benefits from the DP umbrella and reduce harm. DP guarantees to protect only private information (mentioned above). So, if one's secret is general information, it will not be protected!

2.6.3 AoD Techniques for Audio, Video, images etc.

Nowadays, 80 percent of data captured is audio, video and image format. AOD for multimedia is not easy and it may be a combination of different techniques. You can use one technique or group of techniques to anonymise multimedia data. List of techniques used as mentioned below:

2.6.3.1 Cryptographic methods

2.6.3.1.1 Homographic encryption

Homomorphic encryption is a form of randomized encryption. When employed as part of a de-identification technique, homomorphic encryption is able to be used to replace any identifying or sensitive attribute within a data record with an encrypted value.

2.6.3.1.2 Homomorphic Secret sharing

Homomorphic secret sharing enables a secret to be divided into "shares", specified subsets of which are used to reconstruct the secret, such that if the same mathematical operation is performed on all the shares used to

reconstruct the secret then the result is the effect of performing this mathematical operation on the original secret. When employed as part of a de-identification technique, homomorphic secret sharing can be used to replace any identifying or sensitive attribute within a data record with two or more shares produced by a message sharing algorithm. These shares can then be distributed to two or more shareholders, the number of which is determined by the instantiation of the secret sharing scheme.

2.6.3.1.3 Order-preserving encryption

It is a form of non-randomized symmetric encryption. When employed as part of a de-identification technique, order-preserving encryption can be used to replace any identifying or sensitive attribute within a data record with an encrypted value.

2.6.3.2 Non-cryptographic methods

2.6.3.2.1 K-anonymity, based on generalization

K-anonymity is a formal privacy measurement model that ensures that for each identifier there is a corresponding equivalence class containing at least K records. While the resulting dataset has limited (i.e. 1/K) link ability, it does not contain measures designed to prevent potential inference attempts.

2.6.3.2.2 L-diversity

L-diversity is an enhancement to K-anonymity for datasets with poor attribute variability. It is designed to protect against deterministic inference attempts by ensuring that each equivalence class has at least L well-represented values for each sensitive attribute. L-diversity is not a single model but a group of models. Each model has diversity defined slightly differently, e.g. by counting distinct values or by entropy.

2.6.3.2.3 T-closeness

T-closeness is an enhancement to L-diversity for datasets with attributes that are unevenly distributed, belong to a small range of values, or are categorical. It is designed to protect against statistical inference attempts, as it ensures that the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the attribute in the overall dataset is less than a threshold T. This technique is useful when it is important for the resulting dataset to remain as close as possible to the original one.

2.6.3.2.4 Permutation

De-identification technique for reordering the values of a selected attribute across the records in a dataset without modifying these values

2.6.3.2.5 Masking

The term “masking” refers to a de-identification technique that involves removing all direct identifiers from the dataset and potentially stripping out some or all of the additional remaining identifying attributes for all records in the dataset. Removing a portion of a direct identifier so that it is no longer a unique identifier is also considered to be a masking technique.

2.6.3.2.6 Differential Privacy

Formal privacy measurement model that ensures that the probability distribution of the output from a statistical analysis differs by at most a specified value, whether or not any particular data principal is represented in the input dataset [14]

- Examples of Differential Privacy
- Reasons for Differential Privacy

DP has valuable properties that makes it become a rich framework for analysing sensitive personal information and privacy protection:

2.6.3.2.6.1 Quantification of privacy loss

Privacy loss is a measure in any DP mechanisms and algorithms. It permits comparisons among different techniques. Privacy loss is controllable that ensures a trade-off between it and the accuracy of general information.

2.6.3.2.6.2 Composition

The quantification of loss permits the analysis and control of cumulative privacy loss over multiple computations. Understanding the behaviour of differentially private mechanisms under composition enables the design and analysis of complex differentially private algorithms from simpler differentially private building blocks.

2.6.3.2.6.3 Group Privacy

DP permits the analysis and control of privacy loss incurred by groups, such as families.

2.6.3.2.6.4 Closure Under Post-Processing

DP is immune to post-processing: A data analyst, without additional knowledge about the private database, cannot compute a function of the output of a differentially private algorithm and make it less differentially private.

Differential privacy can be used to quantify privacy. It's important to remember that privacy guarantees deteriorate with repeated use, so it's worth thinking about how to mitigate this, whether that be with privacy budgeting or other strategies.

2.7 Criteria for Selection of Anonymization techniques

Organizations can select the anonymization techniques as per the applicability according to the various factors such as business requirements, security considerations and regulatory requirements.

Also, they may choose a hybrid model of applying more than one anonymization technique on the data to achieve the objective of eliminating the identifiers. Following table derives the sample criteria for selection of the anonymization techniques:

Technique Name	**Criteria for selection of Anonymisation techniques**
Attribute Suppression	When an attribute is not required in the anonymised dataset.

Technique Name	Criteria for selection of Anonymisation techniques
Record Suppression	When only specific **record**s are not required in the anonymised dataset
Character Masking	When the data value is a string of characters and hiding part of it is sufficient to provide anonymity required
Pseudonymisation	When data values need to be uniquely distinguished and when the anonymised data is required to be irreversible (where the original values are disposed) and reversible (where the identity database is securely kept and not shared).
Generalisation	When the anonymised data is required to make inferences about broader trends or patterns.
Swapping	When the anonymised data set did not require to match the initial information and the parameter values required to remain the same.
Data perturbation	When the anonymized dataset is required to allow users to ascertain key summary information about the data that is not distorted.
Synthetic data	When the anonymised data set requires no connection to the real data set by replacing the real data with the artificial data.
Data aggregation	When the anonymised data set requires to be in summary form for statistical analysis by aggregating the data and consolidation of the same as per the requirements.

2.8 Case Study

2.8.1 AoD Techniques for specialized Data

ImdpGAN (information maximizing differentially private Generative Adversarial Network) is an end-to-end framework that simultaneously achieves privacy protection and learns latent representations. Below figure shows the application of imdpGAN framework to CelebA dataset to show how the privacy and learned representations can be used to control the specificity of the output. The CelebA dataset includes 202,599 celebrity face images with variations like pose and brightness.

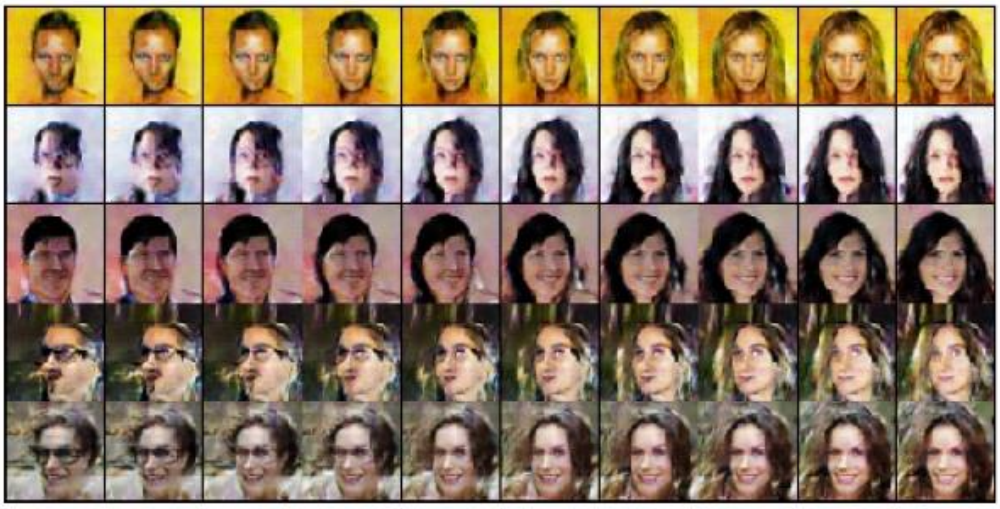


Fig2.a: Changing hairstyle

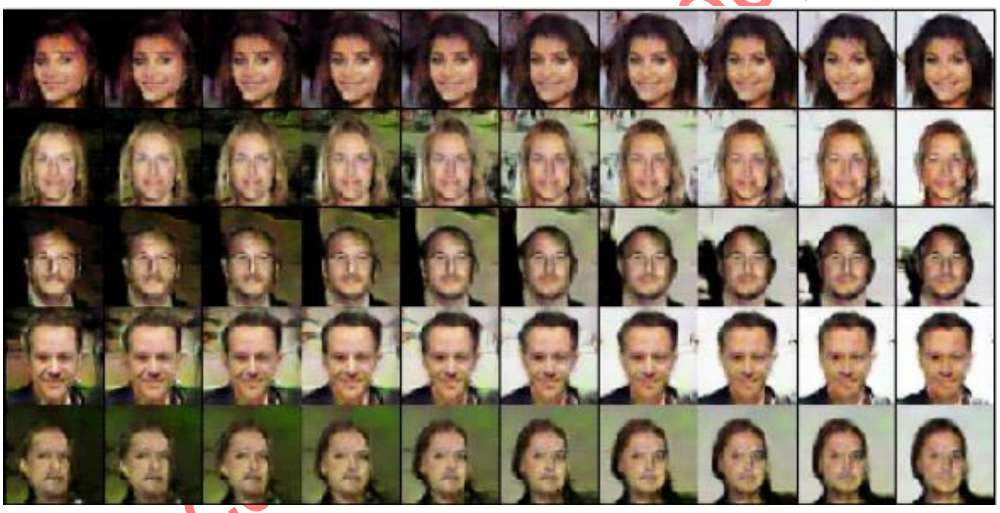


Fig2.b: Changing brightness

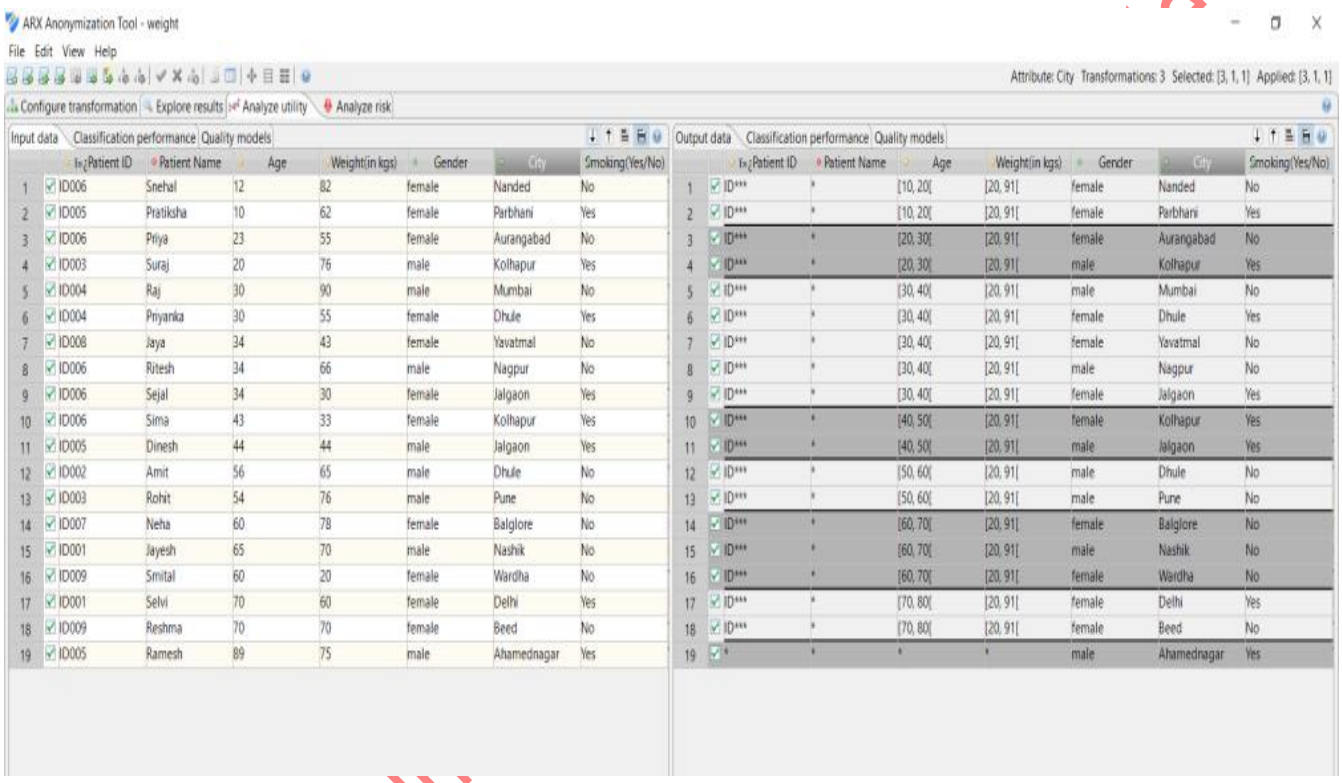
2.8.2 AoD Techniques for regular data set

Below is one of the sample case studies to depict the usage of anonymization techniques for regular dataset for easy understanding and reference.

Hospital ABC maintains various medical records such as Patient ID, patient name, age, weight, gender etc., There was a requirement to share the patient's record for analysis purposes to XYZ. Hospital team performs the anonymization on the data sets of patient records using the open source tool called “ARX” and then shares the anonymized data sets to the XYZ. Applicability of anonymization techniques is dependent on case-to-case basis and subjective in nature. The anonymization techniques used in this case study are listed below:

- Record suppression technique was applied on the Patient name, considering the name of patient was not required for the analysis.
- Generalization techniques were applied on the age, and weight of the patients to de-identify the individual to enhance the data privacy.

Below is the screenshot of data before anonymization on the left side and after anonymization on the right side.



The steps in which they have performed the anonymization of patient records using ARX tool are mentioned in the Annexure-4: Case Study. *Please note that the open-source tool ARX mentioned here for explanatory purposes only and not to be considered as recommendation.*

Chapter 3: Stakeholders, Governance, Audit & Feedback Mechanism

This chapter dwells upon the implementation aspects of AoD and covers identifying various stakeholders, framing Standard Operating Procedures (SOPs) and putting in place appropriate governance, audit and feedback mechanisms that may be integrated into implementing organisations PIMS framework. It also covers aspects of capacity building and creating awareness about the anonymisation of data.

3.1 Stakeholders

All those who capture, process, store, share or use the data are stakeholders of data anonymisation. The list of these stakeholders is as below:

3.1.1 Professional Users (Users of the data captured/ processed by any e-governance organisation)

Many applications need data for analysis and research purposes. Such data is captured by an owner application and then shared with these professional users. An application that captures the data is referred to as an "Owner Application". The owner application should wherever feasible share only anonymised data with user applications.

Users can be of different types

- ***Application Users*** – Departments or applications using or integrated with the source application have to access the data generated by the source application. In this scenario, data sharing is happening between two systems.
- ***Citizens*** – The users of the owner application also access the data from the owner application online (on-screen information) or offline (such as in print) format.
- ***Call centres and Third-party Service providers*** - Automated verification of the people's information even without third-party service providers having access to sensitive data of the source application.
- ***Researchers or data analysers*** - The users who need the data for analysis and research purposes and expect that the data provided to them is de-identified.

3.1.2 Processors

The owner application captures Personal Identifiable Information (PII) or sensitive information. The application then processes the information in order to achieve different objectives. Teams involved in processing the data are referred to as "processors". These teams are responsible for converting raw personal data into anonymised form by various technical procedures/ tools detailed in the previous chapter. The processing organisation includes partners/ contractors, various teams associated with the Source Application such as the development team, testing team, Production support, system admin, infrastructure, etc., who have direct access to the captured data.

3.1.3 Auditors and Reviewers

As AoD is crucial to preserve an individual's privacy and make data more secure, it needs to be audited and reviewed. Auditors are responsible for assessing processed data to identify whether an individual is identifiable by combining de-identified or anonymised data with other data using both direct and indirect methods. The auditors and reviewers can be Compliance Officer, Legal staff, External Auditors/ Reviewers. They are important stakeholders as they also have access to the data during the audit process.

3.1.4 Data Principal

Data Principal are users whose personal data is processed in various e-governance projects for providing services.

3.2 Standard Operating Procedure (SOP)

3.2.1 High-Level Process for de-identification

With the privacy protection bill, it is essential to ensure that the data is anonymised or de-identified. This subsection provides the essential steps to be followed by the Owner organisation or team to ensure that the data is adequately anonymised. In addition, the nodal officer is responsible for taking care of any data (including de-identified data) moving out of the organisation.

- Step 1: Determine dataset that needs de-identification process

The first step is to identify raw data that needs an anonymisation process. The source of raw data may be one or multiple depending on the requirement of the organisation. Data collected from all the sources and in all the applications of the owner organisation should be considered for this step.

- Step 2: Determine the release model/ Policy

Once the data is anonymised or de-identified, it can be shared with business users. In addition, there can be mechanisms such as system to system transfer on request, making the anonymised data available to researchers or data analysts on request.

This step refers to how the de-identified dataset will be released. It may be a Public release or a Limited release. The Public release refers to the release of a dataset publicly without any control. This may pose little challenge on the anonymisation techniques. The limited release refers to the release of a dataset to known research groups in a controlled manner.

- Step 3: Identify roles and responsibilities for overseeing the de-identification process

Data Anonymization is an important activity, and various teams will be involved in it. Hence the owner organisation should prepare a policy document regarding the roles and responsibilities of various teams or individuals involved. A sample of identification roles and their assigned responsibilities is tabulated as follows:

#	Role	Responsibilities
1.	**Chief Investigator (CI)**	Overall responsibility for converting the raw data into anonymised data by following the de-identification process. CI also responsible for delegating responsibilities to other officials
2.	**Co-directors**	Responsible for authorising and overseeing the process for the utilisation of anonymised data
3.	**Trial data manager**	Responsible for the collection and utilisation of data in the agreed manner and accepting and validation of data delegated from the Trial Statistician
4.	**Trial Statistician**	Responsible to CI for all statistical aspects of the trial, including accepting and validating data
5.	**Trial manager**	Responsible for coordinating the arrangements for the appropriate use of data, including ensuring the relevant permissions are in place
6.	**Tester**	Responsible for testing of anonymised data set against the risk of re-identification
7.	**Validator**	Responsible for validation of overall process used in anonymisation
8.	**Auditor**	May be internal or external to the organisation. Responsible for assessing the appropriateness of the methodology used for anonymised data.

Table 3.a: Sample of Roles & Responsibilities Assignment

- ### Step 4: Determine direct identifiers and quasi-identifiers in the dataset

  In order to decide the data to be anonymised and the technique to be used for anonymisation, the data should first be classified as Direct identifiers and Quasi-identifiers.

  Direct identifiers are the fields in datasets that directly disclose the identity of an individual. Example: phone numbers, Aadhaar number, driving license, etc.

  Quasi-identifiers are the fields in datasets that do not directly disclose an individual's identity; however, they may be able to disclose the identity by linking other datasets. Examples: date of birth, age, pin-code, sexual orientation, city, beliefs/ faith, etc.

- ### Step 5: Mask (transform) direct identifiers

  The direct identifiers are masked by using different masking techniques. For example, masking techniques may be (1) removal of the direct identifiers, (2) replacement of the direct identifiers with random values, or (3) replacement of the direct identifiers with pseudonyms.

  Once this step is completed, the database is free from the risk of re-identification from direct identifiers.

- ### Step 6: Perform threat modelling of quasi-identifiers
  Experts should perform threat modelling to identify quasi-identifiers in the data set and what information may get revealed with transformed data (output of step 4).

- ### Step 7: Determine the re-identification risk threshold
  The organisation determines the acceptable risk threshold of re-identification based on the method employed for de-identification. This should be identified based on acceptable risk and controls, based on strong precedents and standards. Accordingly, a report may be created.

- ### Step 8: Determine the transformation process to be used to manipulate the quasi-identifiers
  Determine the transformation process and document the whole transformation process. Techniques recommended in chapter 2 may be used for the transformation process.

- ### Step 9: Import (sample) data from the source database
  Importing data from the source database should be documented as it may be a complex or straightforward exercise depending on the data involved.

- ### Step 10: Review the results of the trial de-identification.
  Perform a trail de-identification based on the steps mentioned above. First, review the process and check if the results achieved are as per the expectations. Next, review the code and algorithms used in de-identification. If errors are there, correct them. The overall risk should be less than re-identification threshold.

- ### Step 11: Transform the quasi-identifiers
  Transform the quasi-identifiers for the entire dataset. Then, based on step 9, use code and algorithms for entire datasets.

- ### Step 12: Evaluate the actual re-identification risk
  For every dataset produced in step 10, ask and evaluate “can this information be used to identify someone?”

- ### Step 13: Compare the actual re-identification risk with the threshold specified by the policymakers
  Compare actual re-identification risk with the threshold set by the policymakers. For success, overall risk should be less than reidentification risk threshold.
  If the data do not pass the actual risk threshold, adjust the procedure and repeat Steps 12 and 13.

- ### Step 14: Determine access controls for the data even while sharing it
  Controls include both technical and non-technical (e.g. legal and organisational measures).

- Step 15: Document the whole anonymisation process
  Details of the anonymisation process should be captured in a detailed manner. It will help reviewers and auditors to identify problems in the anonymisation process.

3.2.2 Data Sharing Mechanism

Steps mentioned in section 3.2.1 are to be followed by the owner organisation to anonymise data. Business users who need to use the anonymised data are of two types (1) Researchers and analysts who require such data for their study & (2) Systems or applications that need the data from the owner application as a part of an exchange.

Researchers and data analysts Access control - Nowadays, Businesses acquire a lot of personal data during day-to-day operations. Also, some businesses require de-identified personal data for analysis and research purposes. This section provides the guidelines for businesses for handling and usage of processed personal data.

The owner organisation may anonymise data and make it available in bulk for researchers and analysts for their studies. However, before sharing the data, the owner organisation should take the following information from the researchers and data analysts through a form.

(a) Name of an individual or an organisation with contact details
(b) Which data is required?
(c) Purpose or objective - how the data will be used?
(d) Agreement - shall state that the shared data will not be misused and shared with any third party.

3.2.2.1 High-Level process for obtaining data:

1. To obtain the data, the seeker organisation needs to fill-up the form.
2. The request form should be signed and approved by the Concerned Authority of the data custodian organisation.
3. After receiving the form and the subsequent approval as per Delegation of Power, the Undertaking will be signed between the organisations.
4. After signing the Undertaking, the data will be shared through USB/ E-mail/ SFTP/ approved media.
5. On Completion of the duration of the project, data needs to be expunged by the recipient organisation.
6. The recipient organisation needs to send a confirmation within a specified time limit after the date of expunging the data as declared in the form.

3.2.2.2 High-Level process for a time extension of retaining data:

Seeker organisations may request and get approval from data custodian organisations for a time extension for retaining data.

3.2.2.3 Provision of inspection

During the data usage period, the Auditor may visit the data seeker organisation and check/ confirm whether procedures are being followed as indicated in the signed agreement.

3.2.2.4 Guidelines for Business Users and owner organisation using/ sharing Anonymized Data

Guideline	Remark
Governance	
i. Organisation shall conduct an information audit to determine what information is processed	
ii. Organisation shall conduct an information audit to determine who has access to it	
iii. Organisation shall have a legal jurisdiction for data processing activities	
Data Protection	
i. Organisation shall ensure data anonymised/ de-identified in public domain or otherwise to be irreversible i.e. ● It is not possible to single out an individual ● It is not possible to link records relating to an individual	
ii. Organisation shall ensure it is not possible to have any information for inferencing concerning an individual.	
iii. Organisation shall ensure that the strength of anonymised privacy data will be measured using laid down models to identify re-identification risks.	
iv. Controlled re-identification may be done with organisational and technical measures to make data access to concerned authority/ individual as per relevant legislation.	
v. Organisation shall consider data protection into account whenever processing data	
vi. Organisation shall use de-identification techniques wherever possible	

Guideline	Remark
vii. Organisation shall implement internal security policy for their employees	
viii. Organisation shall conduct awareness trainings for their employees who access personal data as a part of operational security	
ix. Organisation shall conduct periodic assessment of personal data	
x. Organisations shall have an incident management plan for unfortunate incidents such as data breach, exposure of personal data etc.	
xi. Organisation shall inform supervisory authority in its jurisdiction within 72 hours in case of an unfortunate incident such as data breach, exposure of personal data etc. as per relevant legislation.	
Accountability	
i. Organisation shall designate a person for ensuring and implementing privacy laws	
ii. Organisation shall sign strict agreement related to data protection with third party who processes personal data on organisation's behalf	
Privacy Rights	
i. Organisation shall privilege its customers to see what personal data it has about them and how it is being used	
ii. Organisation shall privilege its customers to update their personal data	
iii. Organisation shall privilege its customers to take request for deleting their personal data	
iv. Organisation shall privilege its customers to take request for stopping the processing of their personal data	

3.3 Governance (Monitoring and Compliance)

This section covers guidelines related to different governance aspects.

3.3.1 Recommendations for the Owner Organisation

- The respective departments will review/ internal audit the data anonymisation activities periodically at least once a year, or earlier as and when required due to change in the requirement or implementation/ complaint(s)/ incident(s).
- The data anonymisation process will be made part of the IT Software Development and Maintenance process.
- The project should be audited by a third-party annually for compliance to Data Privacy, including data anonymisation, to ensure the effectiveness and correctness of the data anonymisation.
- Preferably, use tools to test if the data anonymisation has been completed instead of only manually verifying it.
- Employ tools for performing data anonymisation along with verifying the correctness of data anonymisation.
- Track all the non-conformances to closure through effective corrective action based on risk assessment to prevent a recurrence.
- Third-party vendors and their subcontractors should also be in the audit scope as they do have access to the data.

3.3.2 Checklist/ guidelines for review/ internal audit

A sample checklist for reference of owner organisations for data anonymisation is as follows. The auditing agency can further expand the checklist. The sample checklist covers minimum and essential points to be followed by the owner organisation.

Guideline	**Remark**
Governance	
i. Third-party information audit to be conducted. Various policy documents created, data identified for anonymisation and de-identification, process defined, data sharing policy, anonymised data- all these should be provided to the auditors for verification.	
ii. Organization should have an access control policy. The same should be audited.	
iii. The organisation should have a policy regarding data anonymisation which should be compliant with law for the purpose. The policy should be approved by senior management and vetted on a regular basis.	

Guideline	Remark
Process Related	
i. List techniques and algorithms used for de-identification	
ii. List software(s) used for de-identification	
iii. Document the qualification of individuals who performed de-identification	
iv. Provide the possible tests to determine effectiveness of de-identification	
v. Details of the access controls applied for usage of de-identified data	
vi. List of direct identifiers identified with raw dataset with justification of exclusion if any.	
vii. List of quasi-identifiers if direct identifiers are not available with justification	
viii. De-identification process policy and where it is used	
ix. Information about original raw dataset status (retained or deleted) and the access controls for the whole process.	
x. Justification document regarding decisions made for the whole process of de-identification and re-identification	
xi. Information about whether re-identification is possible? If yes, what controls applied for mitigation?	
xii. List mechanism available to perform any successful re-identification attempt of datasets	
xiii. Document problems if the data set is re-identified	

Guideline	Remark
xiv. List of de-identified data sets published in public domain or shared with other organisations for processing	
xv. Details about the risk assessment carried out at regular internal pre and post release of anonymised data	

3.3.3 Perform a risk assessment both pre and post-release of anonymised data

Risk Management comes in handy to supplement the imperfections of anonymisation. Given the variety of harms that can result from the use or distribution of de-identified data, organisations should focus their efforts on risk minimisation throughout the process when concerning use of anonymised data sets. Risk assessment should determine the threats and vulnerabilities and derive acceptable risk appetite. The higher the potential negative impact, the lower the risk threshold should be. Depending on quantum and sensitivity of personal information, storage-processing of identified data sets, what stage of information lifecycle is anonymisation being done, anonymity technique used, disclosure and access by number of parties, use of resultant anonymised data sets etc. risk assessment will vary. Appropriate controls should be put in place to mitigate risk throughout the process. Following the methodology, once risk assessment is done and data anonymisation proceeds, another risk assessment should follow to check if acceptable risk levels are being implemented in the process. Similarly, risk assessment may be done post release of data considering new vulnerabilities and threats surfaced and appropriate technical and procedural controls may be applied on sending the anonymized data set to the third party/within the organization. The importance of risk identification is also significant as unacceptable risk of identification of data principals would bring the data processing back under the scope of upcoming data protection law.

There are no definite rules as to what risk threshold values should be as these are case dependent. In context of risk mitigation in determining identifiability/ link-ability/ inferencing from anonymised data sets, some re-identification attack scenarios such as

- Unsanctioned access and use of data
- Disclosure of non-anonymised and anonymised data due to lack of awareness
- Insider attack with privilege access to process data
- Known risk scenarios such as Prosecutor risk/ Journalistic risk/ marketer risks
- Data Sharing/ Publishing/ Disclosure/ Breach: Even with anonymised data, any use case may give rise to a loss of privacy.

Even on part of organisations receiving anonymised data for processing, they should agree not to attempt reidentification (unless explicitly approved) and provide similar level of re-identification restrictions as the primary e-governance project, take reasonable steps to keep all involved individuals and entities from reidentifying data, and keeping anonymised data confidential (unless required otherwise)

3.4 Data Privacy Incident Reporting

3.4.1 Incident Reporting

All data privacy incidents should be reported to the concerned stakeholders. In addition, a data privacy incident reporting system should be in place for users to report any gaps in data anonymisation or any other personal information breach. Time to time, laws and regulatory guidance are issued to specify on the format, agencies and users to be intimated on data breach.

3.4.2 SLAs

Data Incidents should be acknowledged within a specified time frame as per limit and resolved based on the severity.

3.4.3 Audit

During Internal Audits, Data Privacy Incidents should be reviewed.

3.5 Capacity Building and Awareness Creation

Capacity building and creating awareness are essential for the successful implementation of anonymisation of data in various projects. In this section, aspects of human resource development, awareness programs are deliberated.

3.5.1 Competency Development (People)

People involved in e-Governance projects need to be trained on data privacy controls and anonymisation. In addition, they need to understand how to handle data across its lifecycle. Typically, the data anonymisation techniques awareness program will be a part of the Data Privacy Awareness Program.

3.5.2 Privacy awareness training and assessment

The following users should mandatorily complete privacy awareness training and assessment.

3.5.2.1 Target Audience

- Project owners and management
- IT, Network and Data centre heads
- IT Users
- Auditors and Reviews (development team, testing team, infrastructure team)
- Data Consumers (Other Applications team)

3.5.2.2 Training contents

- Laws, Regulations, Standards and Guidelines
- Data lifecycle
- Personal data identification
- Data Anonymisation techniques
- Risk Management

- Using anonymised data
- Data archival
- Handling data breach

3.5.2.3 Training modes

- In-person
- Virtual
- Web-based

3.5.2.4 Periodic assessment

Assess the knowledge of participants on training content every year.

3.5.2.5 Recordkeeping

Maintain records of the people trained on Data privacy and Anonymisation. The records should include:

- Name
- Employee Number
- Date of examination
- Percentage of marks
- Date of re-examination

3.5.2.6 Awareness

Create data privacy awareness amongst citizens using posters and media advertisements to raise awareness about personal data protection.

Annexure 1: Terms of Reference

The Terms of Reference (ToR) of the Working Group on "Anonymisation of Data" are as follows:

i. To study and review various available national & international standards/ guidelines/ frameworks or other relevant material related to the Anonymization of Data;
ii. To study the work of international organisations and their committees related to the areas of work of WG and formulate the Guidelines for Anonymization of data specific to India;
iii. To constitute sub-committees (if required) for the sub-areas under the main theme area, define its scope and coordinate their activities, as per requirements;
iv. To study comments/ change requests (if any) and re-draft the document as per requirements;
v. To work on suggestions/ revisions within stipulated time period in case of public feedback or observations from ministries/ states/ UTs;
vi. To ensure that the finalized guidelines are adaptable, fair, easy to use;
vii. To submit the finalized implementable & adoptable guidelines/standards on Anonymization of Data along with advice and recommendation within the stipulated time; and
viii. To maintain secrecy regarding draft versions and same to be strictly circulated only among the WG members until the draft / final versions are approved; public distribution is to be strictly prohibited without permission of C-DAC/STQC Directorate/MeitY.

Annexure 2: List of contributors

Working group was constituted by C-DAC, Pune, under the leadership of **Shri Avinash Agarwal**, Director (IT), Telecommunication Engineering Centre (TEC), Department of Telecommunications (DoT), Government of India. The group represented an eclectic collection of professionals. Following were the contributors during the development of this document:

- **Shri Avinash Agarwal**, Deputy Director General (Convergence & Broadcasting), TEC, DOT, Delhi
- **Shri Ankit Jain**, Scientist 'C', STQC ERTL(N), Delhi
- **Prof Arun Balaji Buduru**, Assistant Professor (CSE) & Head - Center of Technology in Policing, IIIT-Delhi
- **Shri Chittaranjan Das**, Scientist 'G', ERTL(E), STQC, Kolkata
- **Dr Mahesh Kuruba**, Advisor – AIOps Strategy and Transformation Client Services, Digitate
- **Dr Mangesh Gharote**, Scientist, TCS Research & Innovation
- **Shri Manoj Pandian A**, Tech. Risk Manager, ReBIT, Navi Mumbai
- **Dr Padmaja Joshi**, Senior Director, C-DAC, Mumbai
- **Shri Rahul Sharma**, Founder, The Perspective
- **Dr Sachin P Lodha**, Head of Cybersecurity and Privacy Research, TCS
- **Shri Sandeep Pandey**, AVP (GRC), Goods & Services Tax Network
- **Ms Shilpa Oswal**, Principal Technical Officer, C-DAC, Mumbai
- **Shri Shubhanshu Gupta**, Principal Technical Officer, C-DAC, Pune
- **Shri Srinivas Poosarla**, Senior VP, Chief Privacy Officer & DPO, Infosys

Annexure 3: Data Anonymization Tools

There is a wide range of open-source data anonymization tools available. Following is the list of some open-source anonymization tools. The selection should be carefully made according to the purpose for which organisation wants to anonymize data. (The list is not in any particular order)

1. ARX Data Anonymization Tool

The ARX Data Anonymization Tool is an open source and cross-platform tool. It supports a wide variety of privacy and risk models. ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface.

ARX is also available as a comprehensive software library with a clean API that delivers data anonymization capabilities to any Java program.

ARX supports arbitrary combinations of the following data transformation models

- Global and local transformation schemes: ARX can apply the same transformation scheme to all records in a dataset or apply different transformation schemes to different subsets of records.
- Random sampling: Privacy risks can be reduced by drawing a random sample from the input dataset.
- Generalization: Records can be made less unique by generalizing attribute values based on user-specified hierarchies.
- Record, attribute and cell suppression: Privacy risks can be lowered by removing individual attribute values or complete records.
- Micro aggregation: Clusters of numeric attribute values can be combined into a common value by user-specified aggregation functions.
- Top- and bottom-coding: Values exceeding a user-defined range can be truncated.

Ref: https://arx.deidentifier.org/

2. Amnesia

Amnesia is a flexible data anonymization tool that allows to remove identifiable information from data sets. Amnesia does not only remove direct identifiers like names, Aadhaar Number, etc., but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data. Amnesia supports k-anonymity and km-anonymity.

Amnesia is implemented in java and javascript and it can be used as a standalone application or as a service. Moreover, it provides a ReST service API to allow the incorporation of its anonymization engine to other systems.

Ref: https://amnesia.openaire.eu/

3. µ-ARGUS

µ-ARGUS is a tool designed to create safe micro-data files and is based on the programming language R, which is specifically built to support statistical analyses. ARGUS stands for 'Anti Re-identification General Utility System'. The tool uses a wide range of different statistical anonymization methods such as global recoding (grouping of categories), local suppression, randomisation, adding noise, micro-aggregation, top- and bottom coding. It can also be used to generate synthetic data.

4. SDCMicro (Statistical disclosure control)

SDCMicro is a free, R-based open-source package for the generation of protected microdata for researchers and public use. Data from statistical agencies and other institutions are mostly confidential. This package can be used for the generation of anonymized (micro)data, i.e. for the creation of public- and scientific-use files. In addition, various risk estimation methods are included. The associated package sdcMicroGUI includes a graphical user interface for various methods in the sdcMicro package.

Ref: https://cran.r-project.org/web/packages/sdcMicro/index.html

5. Anonimatron

Anonimatron is a tool that pseudonymizes datasets and that can be used to generate pseudonymized production data to find a bug or do performance tests outside of the client's production environment. With release of the GDPR, a feature was added that enables the anonymization of files.

Features:

- Anonymize data in databases and files.
- Generates fake email addresses, fake Roman names, and UUID's out of the box.
- Easy to configure, automatically generates example config files.
- Anonymized data is consistent between runs. No need to re-write your tests to handle random data.
- Extendable, easily implement and add your own anonymization handlers
- Multi-platform, runs on Windows, Mac OSX, Linux derivatives.
- Multi database, uses SQL92 standards and supports Oracle, PostgreSQL and MySQL

Ref: https://realrolfje.github.io/anonimatron

6. NLM (National Library of Medicine) Scrubber

NLM-Scrubber is a freely available clinical text de-identification tool designed and developed at the National Library of Medicine. The goal of NLM-Scrubber is to produce HIPAA compliant de-identified health information for scientific use; however, the success rate of this goal depends on the input data. It is the data manager's responsibility to evaluate NLM-Scrubber on their datasets and make an informed decision whether NLM-Scrubber is the right tool for their purpose.

Ref: https://scrubber.nlm.nih.gov/

Annexure 4: Case study

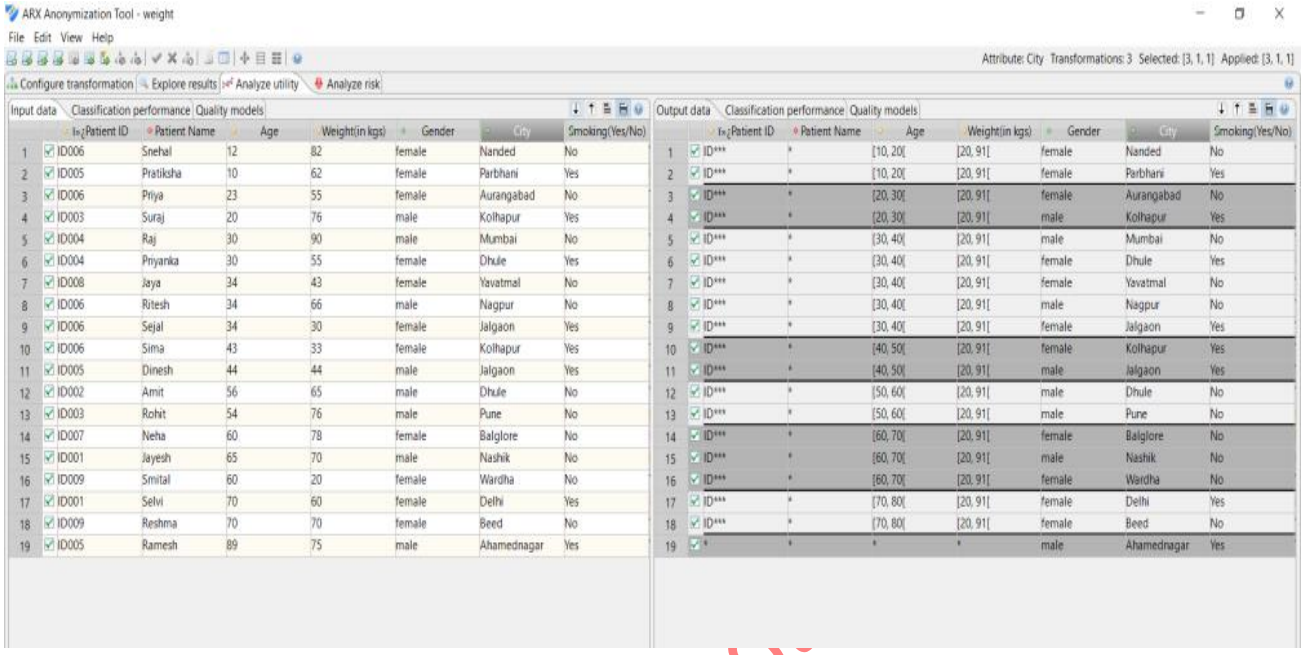
Hospital ABC maintains various medical records such as Patient ID, patient name, age, weight, gender etc., There was a requirement to share the patient's record for analysis purposes to XYZ. Hospital team performs the anonymization on the data sets of patient records using the open source tool called “ARX” and then share the anonymized data sets to the XYZ. Applicability of anonymization techniques is dependent on case-to-case basis and subjective in nature. The anonymization techniques used in this case study are listed below:

- Record suppression technique was applied on the Patient name, considering the name of patient was not required for the analysis.
- Generalization techniques were applied on the age, and weight of the patients to de-identify the individual to enhance the data privacy.

Below is the screenshot of data before anonymization.

Patient ID	Patient Name	Age	Weight(in kgs)	Gender	City	Smoking(Yes/No)	Alchoholic(Yes/No)	Diabetes(Yes/No)	Hypertension(Yes/No)	Thyroid(Yes/No)
ID004	Raj	30	90	male	Mumbai	No	Yes	No	Yes	Yes
ID001	Selvi	70	60	female	Delhi	Yes	No	No	No	Yes
ID005	Ramesh	89	75	male	Ahamedn	Yes	Yes	Yes	Yes	No
ID006	Priya	23	55	female	Aurangab	No	Yes	Yes	Yes	No
ID007	Neha	60	78	female	Balglore	No	No	Yes	Yes	No
ID009	Reshma	70	70	female	Beed	No	Yes	No	Yes	Yes
ID004	Priyanka	30	55	female	Dhule	Yes	No	No	No	Yes
ID006	Sima	43	33	female	Kolhapur	Yes	Yes	Yes	Yes	No
ID008	Jaya	34	43	female	Yavatmal	No	Yes	Yes	Yes	No
ID001	Jayesh	65	70	male	Nashik	No	No	Yes	Yes	No
ID002	Amit	56	65	male	Dhule	No	Yes	No	Yes	Yes
ID003	Suraj	20	76	male	Kolhapur	Yes	No	No	No	Yes
ID005	Dinesh	44	44	male	Jalgaon	Yes	Yes	Yes	Yes	No
ID006	Ritesh	34	66	male	Nagpur	No	Yes	Yes	Yes	No
ID003	Rohit	54	76	male	Pune	No	No	Yes	Yes	No
ID006	Snehal	12	82	female	Nanded	No	Yes	No	Yes	Yes
ID005	Pratiksha	10	62	female	Parbhani	Yes	No	No	No	Yes
ID006	Sejal	34	30	female	Jalgaon	Yes	Yes	Yes	Yes	No
ID009	Smital	60	20	female	Wardha	No	Yes	Yes	Yes	No

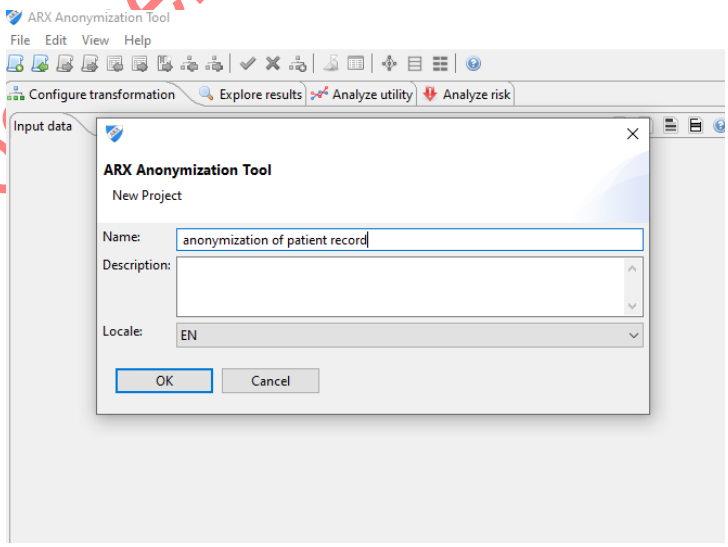
Below is the screenshot of data after anonymization.



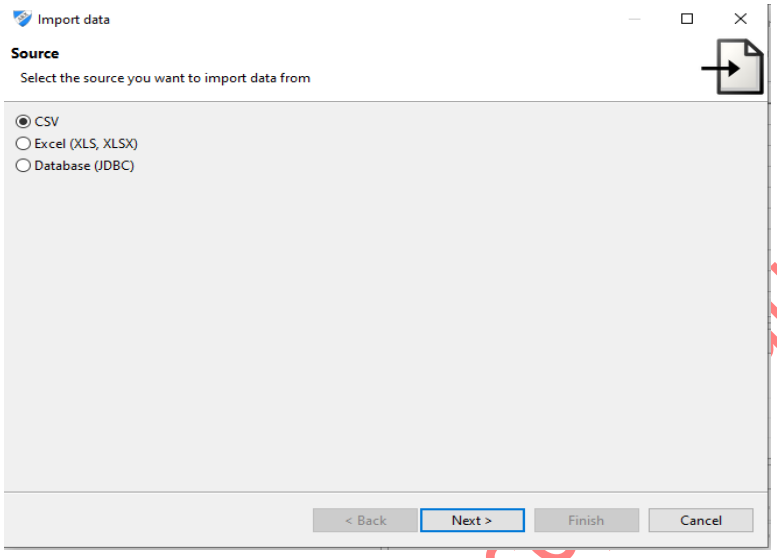
The steps in which they have performed the anonymization of patient records using ARX tool are listed below for reference:

Steps to anonymize data using ARX tool

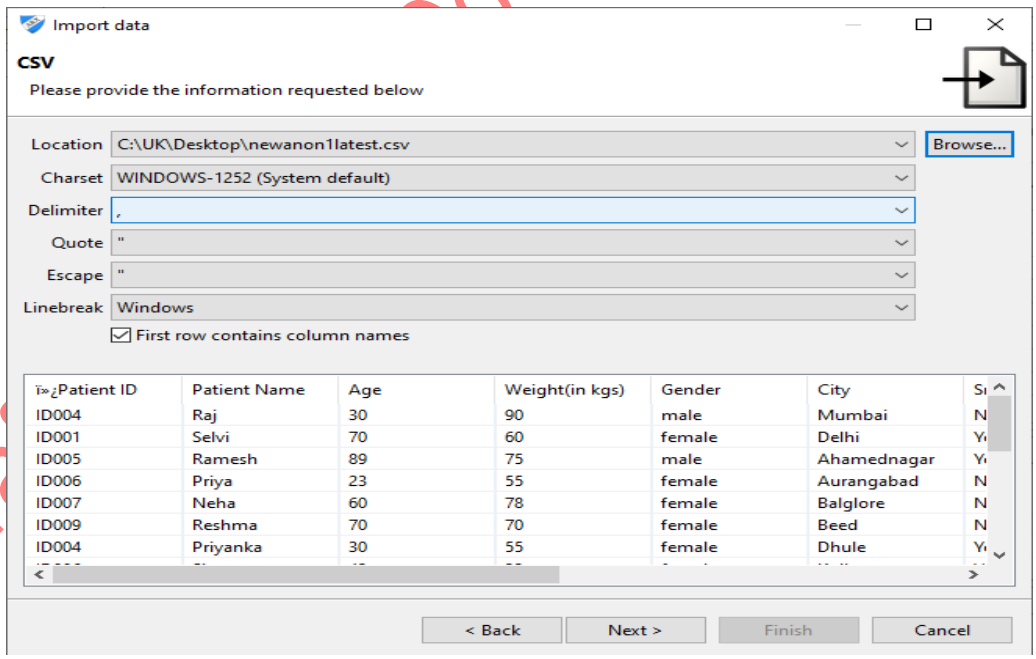
1. Create a new project and give project name.

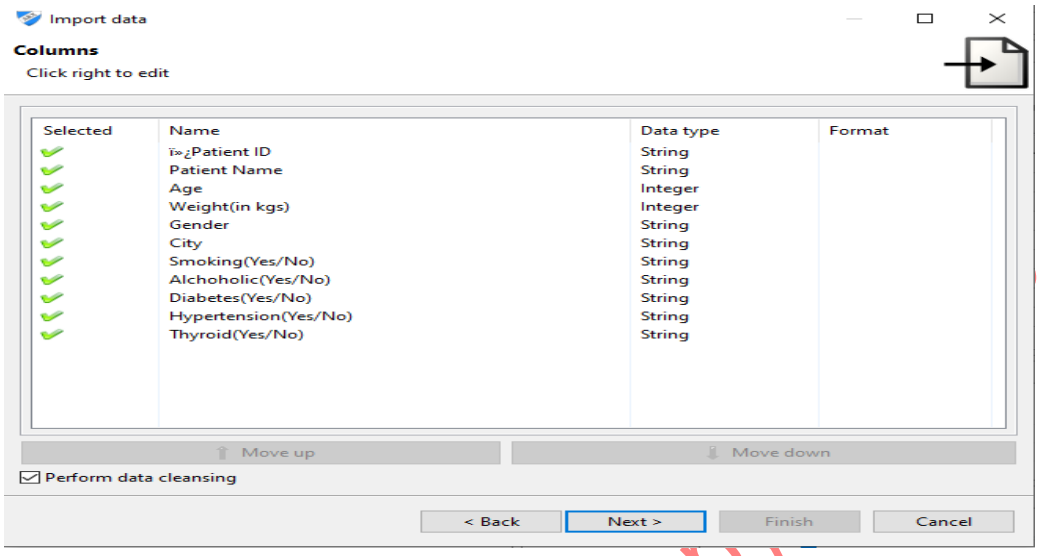


2. Import data from CSV file, Excel or Database.

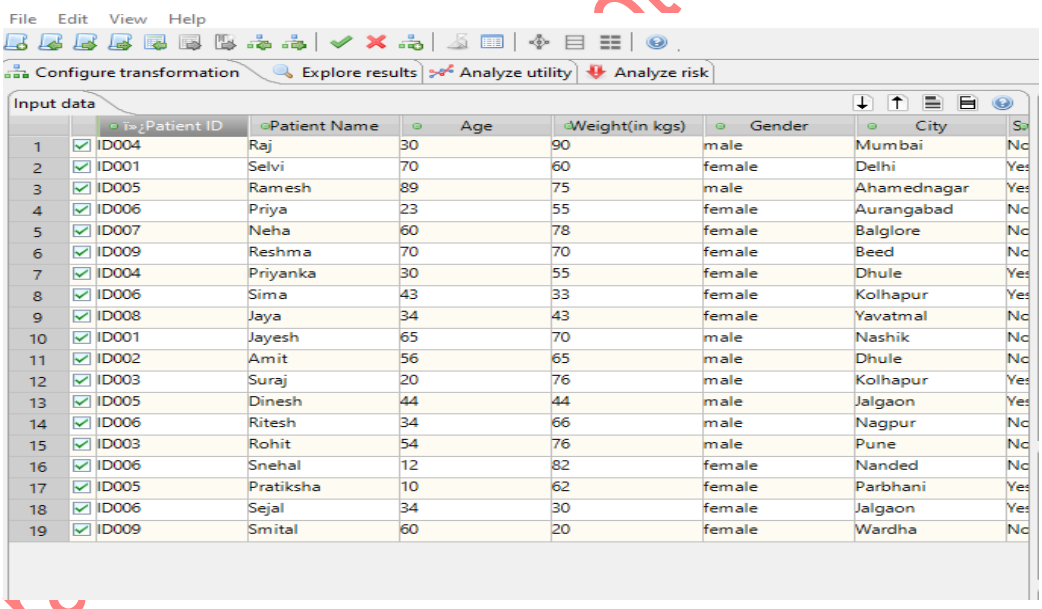


Here the input is taken from CSV file, Browse file from specific location and Click on next. It will display default data types as per the fields in csv file as shown below.



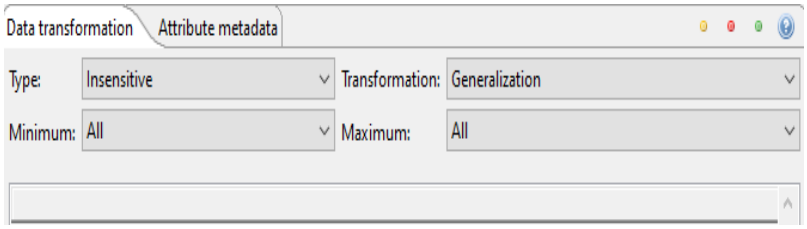


3. Input data is shown on the left hand side.

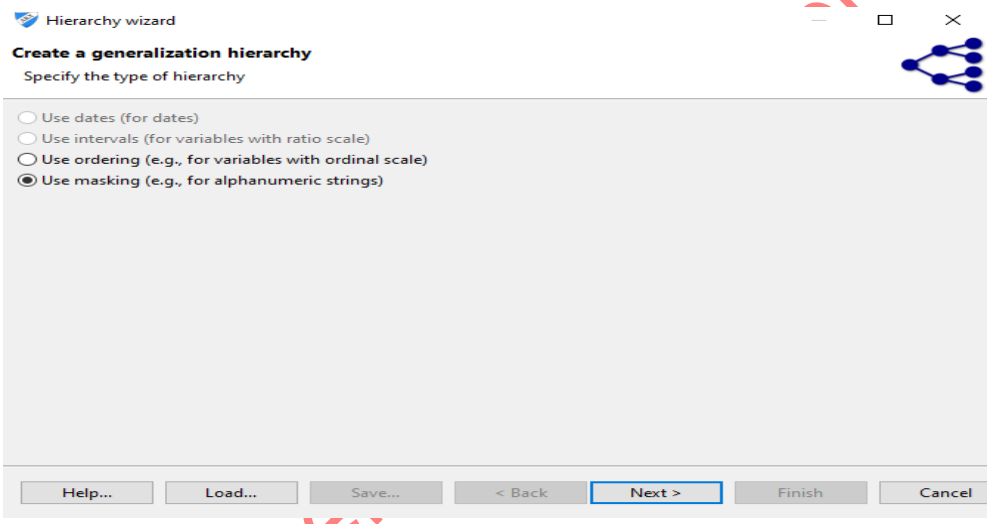


	ï»¿Patient ID	Patient Name	Age	Weight(in kgs)	Gender	City	S
1	ID004	Raj	30	90	male	Mumbai	No
2	ID001	Selvi	70	60	female	Delhi	Yes
3	ID005	Ramesh	89	75	male	Ahamednagar	Yes
4	ID006	Priya	23	55	female	Aurangabad	No
5	ID007	Neha	60	78	female	Balglore	No
6	ID009	Reshma	70	70	female	Beed	No
7	ID004	Priyanka	30	55	female	Dhule	Yes
8	ID006	Sima	43	33	female	Kolhapur	Yes
9	ID008	Jaya	34	43	female	Yavatmal	No
10	ID001	Jayesh	65	70	male	Nashik	No
11	ID002	Amit	56	65	male	Dhule	No
12	ID003	Suraj	20	76	male	Kolhapur	Yes
13	ID005	Dinesh	44	44	male	Jalgaon	Yes
14	ID006	Ritesh	34	66	male	Nagpur	No
15	ID003	Rohit	54	76	male	Pune	No
16	ID006	Snehal	12	82	female	Nanded	No
17	ID005	Pratiksha	10	62	female	Parbhani	Yes
18	ID006	Sejal	34	30	female	Jalgaon	Yes
19	ID009	Smital	60	20	female	Wardha	No

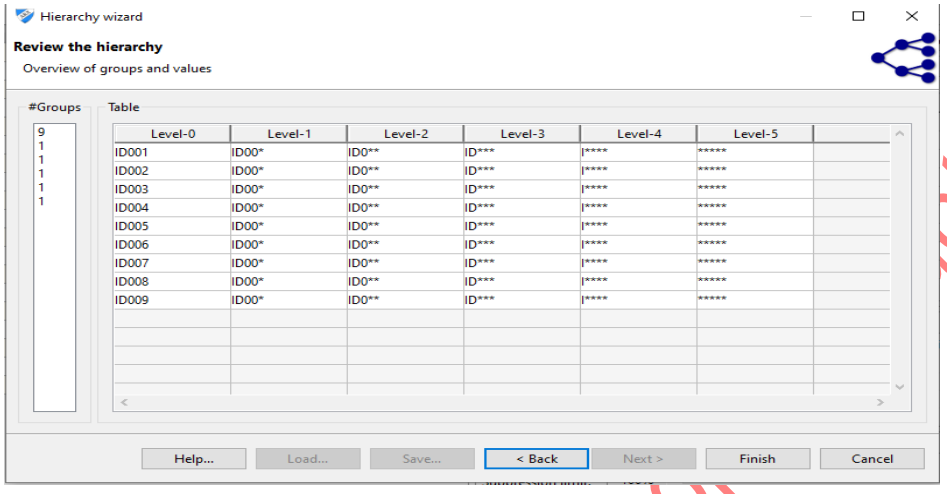
4. Click on attribute and on the right hand side set type as identifying, sensitive, quasi-identifying or insensitive as per data fields in the data set.



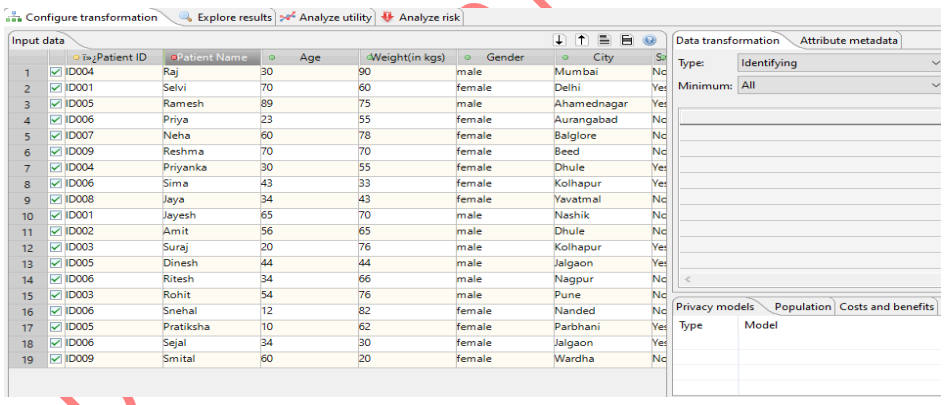
5. Start the anonymization with the Patient ID field. Select quasi-identifying & click on use masking option for masking of patient Id.



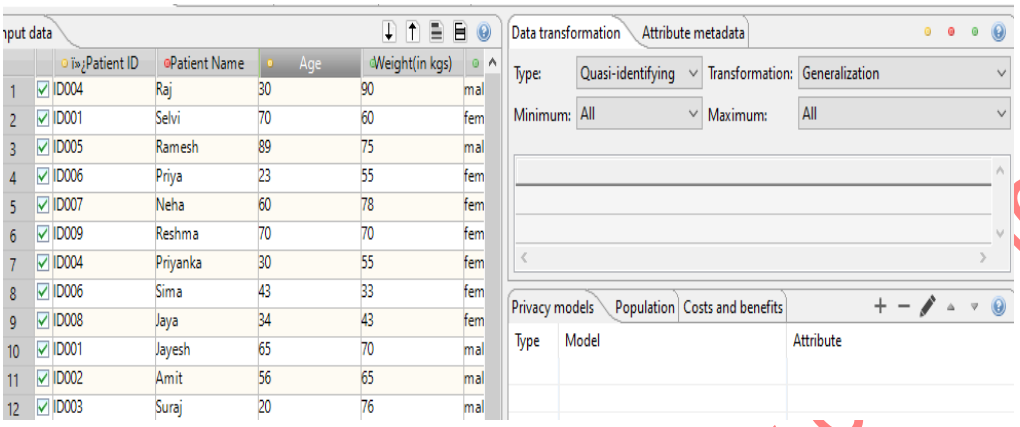
6. Click on finish. Different masking levels are shown on the right hand side, any one out of the same can be selected as per the requirements.



7. As the patient's name is an identifying attribute, select Identifying in type. Identifying columns will be suppressed.

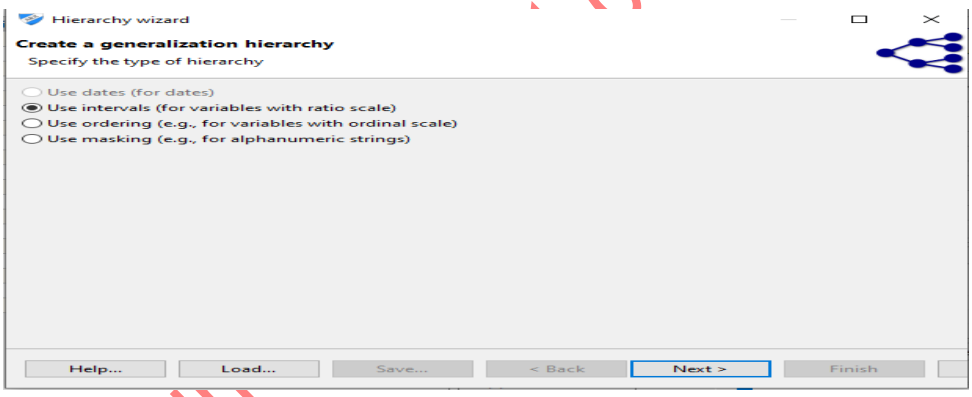


8. Set type for age data as quasi-identifying.

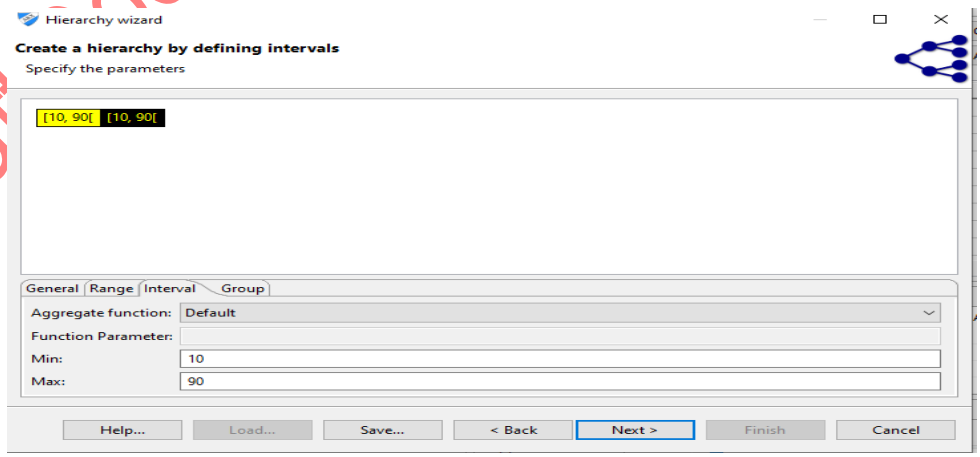


	Patient ID	Patient Name	Age	Weight(in kgs)	
1	ID004	Raj	30	90	mal
2	ID001	Selvi	70	60	fem
3	ID005	Ramesh	89	75	mal
4	ID006	Priya	23	55	fem
5	ID007	Neha	60	78	fem
6	ID009	Reshma	70	70	fem
7	ID004	Priyanka	30	55	fem
8	ID006	Sima	43	33	fem
9	ID008	Jaya	34	43	fem
10	ID001	Jayesh	65	70	mal
11	ID002	Amit	56	65	mal
12	ID003	Suraj	20	76	mal

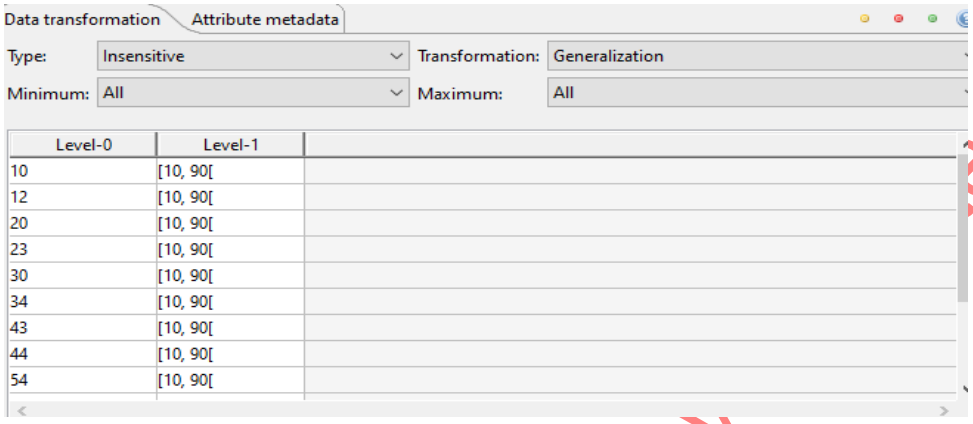
9. To generalize age, set hierarchy for age data and select option of use intervals.



10. Set range for age attribute to generalize it by setting minimum and maximum limit for interval.



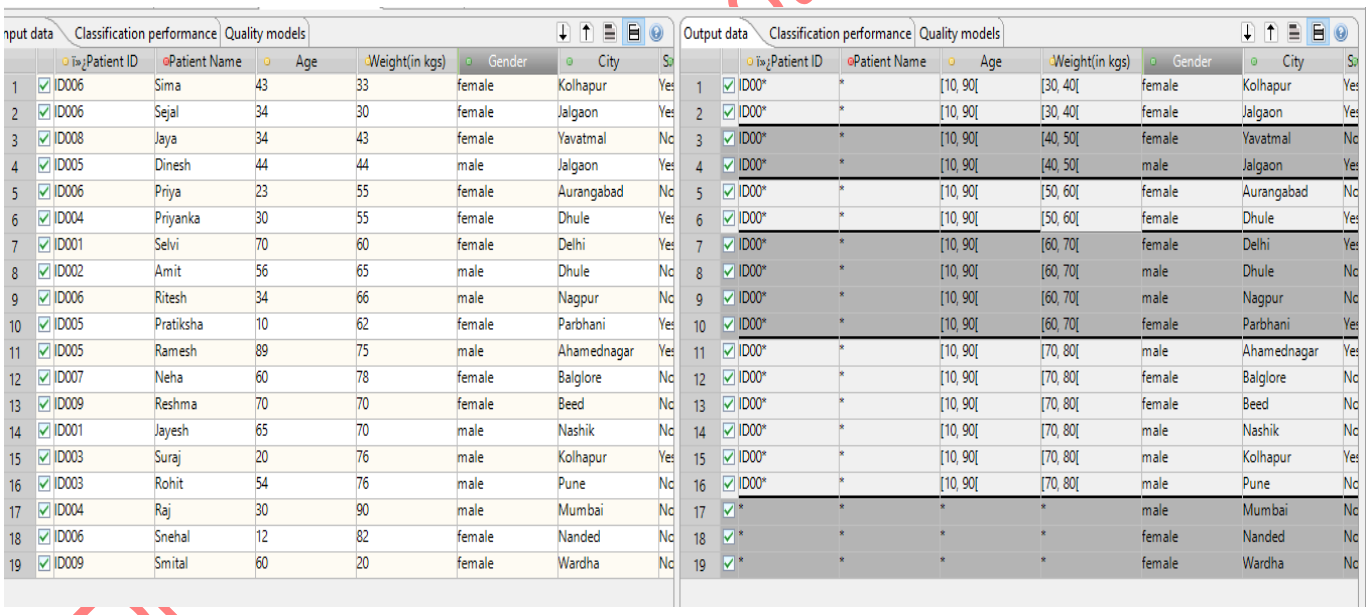
11. Levels are created as shown in below fig, after creating levels set minimum and maximum level as required.



Level-0	Level-1
10	[10, 90[
12	[10, 90[
20	[10, 90[
23	[10, 90[
30	[10, 90[
34	[10, 90[
43	[10, 90[
44	[10, 90[
54	[10, 90[

12. Similar steps like age attribute, are performed for weight attribute.

Once all such settings for all the attributes which are to be anonymized are done, click on analyse utility to view output data. On the right hand side, anonymized data will be displayed.



	Patient ID	Patient Name	Age	Weight(in kgs)	Gender	City
1	ID006	Sima	43	33	female	Kolhapur
2	ID006	Sejal	34	30	female	Jalgaon
3	ID008	Jaya	34	43	female	Yavatmal
4	ID005	Dinesh	44	44	male	Jalgaon
5	ID006	Priya	23	55	female	Aurangabad
6	ID004	Priyanka	30	55	female	Dhule
7	ID001	Selvi	70	60	female	Delhi
8	ID002	Amit	56	65	male	Dhule
9	ID006	Ritesh	34	66	male	Nagpur
10	ID005	Pratiksha	10	62	female	Parbhani
11	ID005	Ramesh	89	75	male	Ahamednagar
12	ID007	Neha	60	78	female	Balglore
13	ID009	Reshma	70	70	female	Beed
14	ID001	Jayesh	65	70	male	Nashik
15	ID003	Suraj	20	76	male	Kolhapur
16	ID003	Rohit	54	76	male	Pune
17	ID004	Raj	30	90	male	Mumbai
18	ID006	Snehal	12	82	female	Nanded
19	ID009	Smital	60	20	female	Wardha

	Patient ID	Patient Name	Age	Weight(in kgs)	Gender	City
1	ID00*	*	[10, 90[	[30, 40[	female	Kolhapur
2	ID00*	*	[10, 90[	[30, 40[	female	Jalgaon
3	ID00*	*	[10, 90[	[40, 50[	female	Yavatmal
4	ID00*	*	[10, 90[	[40, 50[	male	Jalgaon
5	ID00*	*	[10, 90[	[50, 60[	female	Aurangabad
6	ID00*	*	[10, 90[	[50, 60[	female	Dhule
7	ID00*	*	[10, 90[	[60, 70[	female	Delhi
8	ID00*	*	[10, 90[	[60, 70[	male	Dhule
9	ID00*	*	[10, 90[	[60, 70[	male	Nagpur
10	ID00*	*	[10, 90[	[60, 70[	female	Parbhani
11	ID00*	*	[10, 90[	[70, 80[	male	Ahamednagar
12	ID00*	*	[10, 90[	[70, 80[	female	Balglore
13	ID00*	*	[10, 90[	[70, 80[	female	Beed
14	ID00*	*	[10, 90[	[70, 80[	male	Nashik
15	ID00*	*	[10, 90[	[70, 80[	male	Kolhapur
16	ID00*	*	[10, 90[	[70, 80[	male	Pune
17	*	*	*	*	male	Mumbai
18	*	*	*	*	female	Nanded
19	*	*	*	*	female	Wardha

Annexure 5: References and Further Reading

[1] **White Paper of the Committee of Experts on Data Protection Framework for India,** Committee of Experts, Ministry of Electronics and Information Technology (MeitY), 2017
Accessed on 31.08.2021 at
https://www.meity.gov.in/writereaddata/files/white_paper_on_data_protection_in_india_171127_final_v2.pdf

[2] **National Data Sharing and Accessibility Policy**, March, 2012

[3] **The Personal Data Protection Bill**, December, 2019

[4] **ISO 27701-1: Privacy information management — Requirements and guidelines,** August, 2019

[5] **Anonymisation: managing data protection risk code of practice**, Information Commissioner's Office United Kingdom, 2012

[6] **Guide to Basic Data Anonymisation Techniques**, Personal Data Protection Commission Singapore, January, 2018

[7] **GDPR checklist for data controllers**, European Union, 2020
Accessed on 31.08.2021 at https://gdpr.eu/checklist

[8] **Standard Operating Procedures**, Pragmatic Clinical Trials Unit (PCTU), March, 2020
Accessed on 31.08.2021 at https://www.qmul.ac.uk/pctu/media/pragmatic-clinical-trials-pctu/research/PCTU_POL_IG_02-Data-Sharing-Policy-v5.0-(for-publication).pdf

[9] **The Impact of EU Grants for Research and Innovation on Private Firms' Performance**, Gábor Kátay et. al., January, 2019
Accessed on 31.08.2021 at
https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5bf59eb2f&appId=PPGMS

[10] **SOP 23: Standard Operating Procedure for Routine Data Using an Anonymised Data Approach**, West Wales Organisation for Rigorous Trials in Health (WWORTH), May 2014
Accessed on 31.08.2021 at http://www.wales.nhs.uk/sitesplus/documents/1056/WWORTH-SOP23RoutineDataV2.2-140507.pdf

[11] **Standard Operating Procedures for Confidentiality and Protection of Personal Data, UK Clinical Research Collaboration (UKCRC)**
Accessed on 31.08.2021 at https://cdn.ymaws.com/www.tmn.ac.uk/resource/collection/AB6D74CF-ADCA-4A83-9788-A466A380CAD3/UKCRC_CTU_template_SOP_Confidentiality_and_protection_of_personal_data_v1.0.doc

[12] **Probabilistic Anonymity** (pp. 56-79), International Workshop on Privacy, Security, and Trust in KDD, August 2007

[13] **K-anonymity: A model for protecting privacy, International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems**, 10, 557-570, L. Sweeney, 2002.

[14] International Journal of Communication Networks and Information Security (IJCNIS) Vol. 12, No. 1, April, 2020

[15] **imdpGAN: Generating Private and Specific Data with Generative Adversarial Networks**, Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Saurabh Gupta; Arun Balaji Buduru; Ponnurangam Kumaraguru, September, 2020

[16] **Appendix B Concepts and Methods for De-identifying Clinical Trial Data**, April, 2015
Accessed on 31.08.2021 at https://www.ncbi.nlm.nih.gov/books/NBK285994/

[17] **Guidelines for seeking data**, SEBI, October, 2019
Accessed on 31.08.2021 at https://www.sebi.gov.in/reports-and-statistics/statistics/sep-2019/guidelines-for-data-sharing_45488.html