International Standard

ISO/IEC 5207

First edition
2024-04

Information technology — Data usage — Terminology and use cases

Technologies de l'information — Utilisation des données — Terminologie et cas d'utilisation

Reference number
ISO/IEC 5207:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 32, *Data management and interchange*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

The purpose of this document is to provide terminology and use cases in order to support organizations during the decision-making processes that occur throughout the use, sharing and exchange of data.

Given the breadth of data use, exchange and sharing activities, these use cases are presented with a description of the data usage activity including an overview of the data project, objectives, relevant entities involved, and the processes and interventions used in each case.

The use cases are structured to assist users in identifying the decision-making processes within data related activities, irrespective of the business or industry sector context. These use cases can provide users with guidance in considering where control measures can be applied to manage risks within the data process, the data lifecycle or the data environment.

This document can be used in the development of other International Standards and in support of communications among diverse stakeholders and other interested parties.

ISO/IEC 5207 was developed in collaboration with ISO/IEC 5212. Users of this document can refer to ISO/IEC 5212 for additional guidance for the decision-making process for the use, sharing and exchange of data.

Information technology — Data usage — Terminology and use cases

1 Scope

This document sets out terminology and use cases for data use, sharing and exchange. This document provides use cases detailing various types of data usage from both historical and hypothetical perspectives.

This document is applicable to all types of organizations.

2 Normative references

There are no normative references in this document.

3 Terms, definitions and abbreviated terms

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/.

3.1
information
<information processing> knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context has a particular meaning

[SOURCE: ISO/IEC 2382:2015, 2121271, modified — Notes to entry deleted]

3.2
data
re-interpretable representation of *information* (3.1) in a formalized manner suitable for communication, interpretation, or processing

Note 1 to entry: Data can be used for communication, interpretation or processing by humans or automatic means.

Note 2 to entry: Communication, interpretation or processing can include the exchange or sharing of data by one or more entities.

[SOURCE: ISO/IEC 2382:2015, 2121272, modified — Note 1 to entry modified, Note 2 to entry replaced and Note 3 deleted]

3.3
metadata
data (3.2) that defines and describes other data

[SOURCE: ISO/IEC 11179-3:2023, 3.2.30]

3.4
data element
unit of *data* (3.2) that is considered in context to be indivisible

Note 1 to entry: The definition states that a data element is "indivisible" in some contexts. This means that it is possible that a data element considered indivisible in one context (e.g. telephone number) can be divisible in another context, (e.g. country code, area code, local number).

EXAMPLE The data element "age of a person" with values consisting of all combinations of 3 decimal digits.

[SOURCE: ISO/IEC 11179-31:2023, 3.23, modified — domain "<organization of data>" deleted]

3.5
data object
collection of *data* (3.2) that have a natural grouping and may be identified collectively

[SOURCE: ISO/TS 27790:2009, 3.20, modified — "as a complete entity" replaced by "collectively"]

3.6
data type
datatype
named set of distinct values, characterized by properties of those values, and by operations on those values

Note 1 to entry: Images, audio files and video files are considered complex data types.

[SOURCE: ISO/IEC 11404:2007, 3.12, modified — used "data type" instead of "datatype" as preferred term, used "named set" instead of "set" in definition and added Note 1 to entry]

3.7
data set
dataset
identifiable collection of *data* (3.2) available for access or download in one or more formats

Note 1 to entry: A data set can be a smaller grouping of data which, though limited by some constraint such as spatial extent or feature type, is located physically within a larger data set. Theoretically, a data set can be as small as a single feature or feature attribute contained within a larger data set.

Note 2 to entry: A data set can be presented in a tabular form and stored and distributed in tables in word processed documents, spread sheets or databases. It could also be presented in any one of a number of alternative formats, including AVRO, JSON, RDF and XML.

[SOURCE: ISO/IEC 11179-33:2023, 3.5]

3.8
data set collection
curated collection of one or more *data sets* (3.7)

[SOURCE: ISO/IEC 11179-33:2023, 3.6]

3.9
data set distribution
specific available form of a *data set* (3.7) or *data set collection* (3.8)

Note 1 to entry: Each data set might be available in different forms and each of these forms represents a different format of the data set or a different endpoint.

Note 2 to entry: Examples of distributions include a downloadable CSV file, an API or an RSS feed. This represents a general availability of a data set.

[SOURCE: ISO/IEC 11179-33:2023, 3.7]

3.10
data representation
format, *data type* (3.6), character set and other characteristics used to represent *data* (3.2)

Note 1 to entry: *Data types* apply to individual *data elements* (3.4). Formats can apply to sets of data elements, such as records, tables or messages.

3.11
data transformation
conversion from one form of *data representation* (3.10) to another

Note 1 to entry: Transformation of a single *data element* (3.4) can involve a change of character set, *data type* (3.6) or both.

Note 2 to entry: Transformation of a *data set* (3.7) can involve a change of format, such as from XML to JSON, or from a table to a data matrix.

Note 3 to entry: Data transformation does not change the *data* (3.2) value, just the way it is represented. For example, when the letter 'A' is transformed from ASCII to EBCDIC, it is still the letter 'A', just represented in a different character encoding.

Note 4 to entry: Data transformation does not change the semantics of the data.

Note 5 to entry: Where the *metadata* (3.3) for the data includes data provenance, then the metadata should be updated to reflect the data transformation.

3.12
data translation
conversion of a *data* (3.2) value from one code set to another

EXAMPLE 1 Conversion of codes representing human sexes from 'M' or 'F' to '0' or '1', or vice versa.

EXAMPLE 2 Conversion of country codes from 2-alpha code to 3-alpha code or 3-numeric code.

Note 1 to entry: Translation is typically used to convert data from different sources into a standard set of values.

3.13
data product
collection of one or more *data objects* (3.5) that are packaged for or by a specific application

Note 1 to entry: A data product may still provide access to the underlying *data* (3.2) or alternatively be engineered to prevent access to the underlying *data* from which the data product was created.

Note 2 to entry: A data product that has been specifically created to prevent access to the underlying *data* should be noted as such and recorded in the *data set* (3.7) records.

[SOURCE: ISO 21961:2003, 1.5.2, modified — "data items" changed to "data objects". Notes 1 and 2 to entry added]

3.14
data project
programme of work that involves the use, sharing or exchange of *data* (3.2)

3.15
data process
steps involved in the creation, analysis, or alteration of a specific set of *data* (3.2)

3.16
data processing
systematic performance of operations upon *data* (3.2)

[SOURCE: ISO/IEC 23751:2022, 3.8]

3.17
data processing system
computer system
computing system
one or more computers, peripheral equipment, software, human operations, physical processes and *information* (3.1) transfer means that perform *data processing* (3.16)

[SOURCE: ISO/IEC 2382:2015, 2121290, modified — added "human operations and physical processes and *information* transfer means" and notes 1 and 2 to entry deleted.]

3.18
data environment
set of conditions under which *data processing* (3.16) or the *data process* (3.15) occurs

Note 1 to entry: The data environment can include the physical, operational, behavioural and organizational factors which may affect data process outcomes.

3.19
lifecycle
stages involved in the management of an asset

Note 1 to entry: The target of lifecycle in this document is *data* (3.2).

[SOURCE: ISO 55000:2014, 3.2.3, modified — 'life cycle' changed to 'lifecycle', and Note 1 to entry replaced]

3.20
data lifecycle
stages in the management of *data* (3.2)

Note 1 to entry: The target of *lifecycle* (3.19) in this document is data.

[SOURCE: ISO/IEC 20547-3:2020, 3.16, modified — "a" deleted from definition, Note 1 to entry reworded]

3.21
party
natural person or legal person, whether or not incorporated, or a group of either

[SOURCE: ISO 27729:2012, 3.1]

3.22
organization
unique framework of authority within which a person or persons act, or are designated to act towards some purpose

Note 1 to entry: The kinds of organizations covered by this document include the following examples:

a) an organization incorporated under law;

b) an unincorporated organization or activity providing goods and/or services including:

   1) partnerships;

   2) social or other non-profit organizations or similar bodies in which ownership or control is vested in a group of individuals;

   3) sole proprietorships;

   4) governmental bodies.

c) groupings of the above types of organizations where there is a need to identify these in information interchange.

[SOURCE: ISO/IEC 6523-1:2023, 3.1]

3.23
entity
party (3.21) or *data processing system* (3.17) with distinct and independent existence from a *data* (3.2) perspective

3.24
data originator
party (3.21) that created the *data* (3.2) and that can have rights

Note 1 to entry: A data originator can be an individual person.

Note 2 to entry: Rights can include the right to publicity, right to display name, right to identity, right to prohibit *data use* (3.30) in a way that offends honourable mention.

[SOURCE: ISO/IEC 23751:2022, 3.2, modified – Note 2 to entry deleted, and Note 3 to entry renumbered as Note 2.]

3.25
data holder
party (3.21) that has legal control over *data* (3.2) to authorize *data processing* (3.16) of the data by other parties

Note 1 to entry: A *data originator* (3.24) can be a data holder.

[SOURCE: ISO/IEC 23751:2022, 3.4, modified — "control to authorize data processing of data" changed to "control over data to authorize data processing of data".]

3.26
data user
party (3.21) that is authorized to perform processing of *data* (3.2) under the legal control of a *data holder* (3.25)

[SOURCE: ISO/IEC 23751:2022, 3.5]

3.27
ratio scale
continuous scale with equal sized scale values and an absolute or natural zero point

[SOURCE: ISO/IEC 23751:2022, 3.11]

3.28
data level objective
DLO
commitment that a *data holder* (3.25) or a *data user* (3.26) makes for a specific, quantitative characteristic of a *data set* (3.7), where the value follows the interval scale or *ratio scale* (3.27)

Note 1 to entry: A data level objective commitment may be expressed as a range.

[SOURCE: ISO/IEC 23751:2022, 3.12]

3.29
data qualitative objective
DQO
commitment that a *data holder* (3.25) or a *data user* (3.26) makes for a specific, qualitative characteristic of a *dataset* (3.7), where the value follows the nominal scale or ordinal scale

Note 1 to entry: A data qualitative objective can be expressed as an enumerated list.

Note 2 to entry: Qualitative characteristics typically require human interpretation.

Note 3 to entry: The ordinal scale allows for existence/non-existence.

[SOURCE: ISO/IEC 23751:2022, 3.13]

3.30
data use
handling or dealing with *data* (3.2) for a specific purpose

Note 1 to entry: This includes reproducing the data but does not include disclosing the data.

[SOURCE: ISO/TS 14265:2011, 2.11, modified — 'information' has been changed to 'data' in both the definition and Note 1 to entry.]

3.31
data exchange
concerning the representation, transmission, reception, storage, and retrieval of *data* (3.2)

[SOURCE: ISO/IEC 20944-1:2013, 3.21.13.1, modified — Note 1 to entry deleted]

3.32
data sharing
access to or processing of the same *data* (3.2) by more than one authorized *entity* (3.23)

Note 1 to entry: Access to or processing of the data can be synchronous or asynchronous.

Note 2 to entry: In this document, data sharing refers to allowing access to, or the execution of operations over, the original *data set* (3.7).

Note 3 to entry: The way in which data are shared fundamentally influences the available controls and the statements needed in a *data sharing agreement* (3.35).

[SOURCE: ISO/IEC 23751:2022, 3.7, modified — 'use of' changed to 'access to' in Note 1 to entry, Note 2 to entry replaced]

3.33
data usage
any activity involving *data* (3.2)

Note 1 to entry: Data usage includes *data use* (3.30), *data sharing* (3.32) and *data exchange* (3.31).

3.34
data usage framework
framework that sets out the characteristics which should be assessed by the *entity* (3.23) in possession of the *data* (3.2) and captured within the *metadata* (3.3) description

3.35
data sharing agreement
DSA
documented agreement that defines, guides and protects *data sharing* (3.32)

Note 1 to entry: A data sharing agreement generally includes a description of *data* (3.2), data sharing scenarios, roles and participants, platforms, processes, requirements and controls, rights, obligations and responsibilities etc.

3.36
data recipient
entity (3.23) that receives *data* (3.2) via *data sharing* (3.32) or *data exchange* (3.31)

3.37
data accountability
accountability for *data* (3.2) and its usage

Note 1 to entry: *Data usage* (3.33) includes *data use* (3.30), *data sharing* (3.32) and *data exchange* (3.31).

[SOURCE: ISO/IEC 38505-1:2017, 3.4, modified — "use" replaced by "usage" and Note 1 to entry replaced.]

3.38
competent person
person who has acquired, through training, qualification, experience or a combination of these, the knowledge and skill enabling that person to correctly perform the required tasks

[SOURCE: ISO 11525-1:2020, 3.4]

3.39
responsible data officer
officially nominated individual with *data accountability* (3.37)

Note 1 to entry: Responsibility should include enterprise-wide governance and utilization of *information* (3.1) as an asset, via *data processing* (3.16), analysis, *data* (3.2) mining, information trading and other means.

Note 2 to entry: The responsible data officer can be an individual which reports to a governing body which oversees data related activities, can be a delegated position for a specific task such as a major financial project or can be a responsibility under a permanent role within an *organization* (3.22) such as Chief Executive Officers (CEOs), Heads of Government Organizations, Chief Financial Officers (CFOs), Chief Operating Officers (COOs), Chief Information Officers (CIOs), or Chief Data Officers (CDOs), and similar roles.

Note 3 to entry: The delegated data authority should be recognized as a *competent person* (3.38).

3.40
chain of custody
demonstrable possession, movement, handling, and location of material from one point in time until another

[SOURCE: ISO/IEC 27050-1:2019, 3.1]

3.41
access level
level of authority required from a resource owner to access a protected resource

Note 1 to entry: In the context of this document, items to which an access level may be specified are limited to a *data set* (3.7), a *data set collection* (3.9) and a *data set distribution* (3.8).

Note 2 to entry: For the public, the level of authority might describe the degree of public availability of a *data set*.

EXAMPLE Public, restricted public and non-public.

[SOURCE: ISO/IEC 11179-33:2023, 3.3]

3.42
confidential information
information (3.1) that is not intended to be made available or disclosed to unauthorized individuals, *entities* (3.23) or *data processes* (3.15)

[SOURCE: ISO/IEC 27002:2022, 3.1.7, modified — “processes” replaced by “data processes”]

3.43
sensitive information
information (3.1) that needs to be protected from unavailability, unauthorized access, modification or public disclosure because of potential adverse effects on an individual, *organization* (3.22), national security or public safety.

[SOURCE: ISO/IEC 27002:2022, 3.1.33]

3.44
identifiable natural person
individual who can be identified, directly or indirectly, in particular by reference to an identification number or one or more factors specific to their physical, physiological, mental, economic, cultural or social identity

[SOURCE: ISO 22857:2013, 3.7, modified — term changed from “identifiable person”, “one” changed to individual”, “his” changed to “their”.]

3.45
personal information
personal data
any *information* (3.1) on or about an identifiable individual that is recorded in any form, including electronically or on paper

EXAMPLE Information about a person's religion, age, financial transactions, medical history, address or blood type.

[SOURCE: ISO/IEC 15944-5:2008, 3.103, modified — added "personal data" as an admitted term. Changed Note 1 to entry to EXAMPLES.]

3.46
data subject
individual about whom *personal data* (3.45) are recorded

[SOURCE: ISO 5127:2017, 3.13.4.01, modified – Note 1 to entry deleted]

3.47
personally identifiable information
PII
any *information* (3.1) that (a) can be used to establish a link between the information and the natural person to whom such information relates, or (b) is or can be directly or indirectly linked to a natural person

Note 1 to entry: The "natural person" in the definition is the *PII principal* (3.48). To determine whether a PII principal is identifiable, account should be taken of all the means which can reasonably be used by the *privacy stakeholder* (3.49) holding the *data* (3.2), or by any other *party* (3.21), to establish the link between the set of PII and the natural person.

[SOURCE: ISO/IEC 29100:2011/Amd1:2018, 2.9, modified — "NOTE" replaced by "Note 1 to entry"]

3.48
PII principal
natural person to whom the *personally identifiable information (PII)* (3.47) relates

Note 1 to entry: Depending on the jurisdiction and the particular data protection and privacy legislation, the synonym *data subject* (3.46) can also be used instead of the term "PII principal".

[SOURCE: ISO/IEC 29100:2011, 2.11]

3.49
privacy stakeholder
natural or legal person, public authority, agency or any other body that can affect, be affected by, or perceive themselves to be affected by a decision or activity related to *personally identifiable information (PII)* (3.47) processing

[SOURCE: ISO/IEC 29100:2011, 2.22]

3.50
PII controller
privacy stakeholder (3.49) (or privacy stakeholders) that determines the purposes and means for processing *personally identifiable information (PII)* (3.47) other than natural persons who use *data* (3.2) for personal purposes

Note 1 to entry: A PII controller sometimes instructs others, e.g. *PII processors* (3.51) to process *PII* on its behalf while the responsibility for the processing remains with the PII controller.

[SOURCE: ISO/IEC 29100:2011, 2.10]

3.51
PII processor
privacy stakeholder (3.49) that processes *personally identifiable information (PII)* (3.47) on behalf of and in accordance with the instructions of a *PII controller* (3.50)

[SOURCE: ISO/IEC 29100:2011, 2.12]

3.52
de-identification
general term for any process of reducing the association between a set of identifying *data* (3.2) and other data about the *data subject* (3.46)

[SOURCE: ISO 25237:2017, 3.20, modified – inserted "other data about"]

3.53
pseudonymization
process applied to *personally identifiable information (PII)* (3.47) which replaces identifying *information* (3.1) with an alias

Note 1 to entry: Pseudonymization can be performed either by the *PII principals* (3.48) themselves or by *PII controllers* (3.50). Pseudonymization can be used by PII principals to consistently use a resource or service without disclosing their identity to this resource or service (or between services), while still being held accountable for that use.

Note 2 to entry: Pseudonymization does not rule out the possibility of there being (a restricted set of) *privacy stakeholders* (3.49) other than the PII controller of the pseudonymized *data* (3.2) which are able to determine the PII principals identity based on the alias and the data linked to it.

[SOURCE: ISO/IEC 29100:2011, 2.24, modified – Notes to entry have been revised]

3.54
data publication
form of *data sharing* (3.32) that makes *data* (3.2) discoverable by any *entity* (3.23)

Note 1 to entry: Data publication may involve an authorized entity who makes the data available for example through publication online.

Note 2 to entry: Data access may still be controlled via mechanisms such as registration and access systems, log in tracking, user identification etc.

Note 3 to entry: Data publication may involve security mechanisms such as *de-identification* (3.52) or *pseudonymization* (3.53) or the use of a *data product* (3.13) which prevents access to the underlying data.

Note 4 to entry: Data publication may be entirely unrestricted creating *public domain data* (3.55) which can persist in perpetuity.

3.55
public domain data
class of *data objects* (3.5) over which nobody holds or can hold copyright or other intellectual property

Note 1 to entry: *Data* (3.2) can be in the public domain in some jurisdictions, while not in others.

Note 2 to entry: The concept of public domain and the difference between this and "publicly available" is subtle and varies between jurisdictions. Readers should make themselves aware of the specific legal requirements that may apply to them.

[SOURCE: ISO/IEC 19944-1:2020, 3.4.4]

3.56
likelihood
probability of something happening

[SOURCE: ISO/IEC/IEEE 15026-3:2015, 3.13]

3.57
consequence
outcome of an event affecting objectives

[SOURCE: ISO/IEC/IEEE 15026-1:2019, 3.4.1]

3.58
data storage
data store
persistent repository for digital *data* (2)

Note 1 to entry: A data store can be accessed by a single *entity* (3.23) or shared by multiple entities via a network or other connection.

[SOURCE: ISO/IEC 20924:2021, 3.1.14]

3.59
non-volatile storage
storage that retains its contents after power is removed

[SOURCE: ISO/IEC 27040:2024,3.2.11]

3.60
storage medium
storage media
material on which digital *data* (3.2) are, or can be, recorded or retrieved

[SOURCE: ISO/IEC 27040:2024, 3.2.16]

3.61
storage device
any component or aggregation of components made up of one or more devices containing *storage media* (3.60), designed and built primarily for the purpose of accessing *non-volatile storage* (3.59)

[SOURCE: ISO/IEC 27040:2024, 3.2.14]

3.62
big data
extensive *datasets* (3.7) – primarily in the *data* (3.2) characteristics of volume, variety, velocity, and/or variability – that require a scalable technology for efficient storage, manipulation, management, and analysis

Note 1 to entry: Big data is commonly used in many different ways, for example as the name of the scalable technology used to handle big data extensive data sets.

[SOURCE: ISO/IEC 20546:2019, 3.1.2]

3.63
cloud computing
paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand

Note 1 to entry: Examples of resources include servers, operating systems, networks, software, applications and storage equipment.

Note 2 to entry: Self-service provisioning refers to the provisioning of resources provided to cloud services (3.1.2) performed by cloud service customers (3.3.2) through automated means.

[SOURCE: ISO/IEC 22123-1:2023, 3.1.1]

3.64
cloud service agreement
documented agreement between the cloud service provider and cloud service customer that governs the covered service(s)

Note 1 to entry: A cloud service agreement can consist of one or more parts recorded in one or more documents.

[SOURCE: ISO/IEC 19086-1:2016, 3.3]

3.65
data broker
party (3.21) that collects *data* (3.2) from one or more sources and sells the data to one or more *data users* (3.26)

Note 1 to entry: In the context of data broker, sell means to provide data in exchange for money or other item of value.

[SOURCE: ISO/IEC 23751:2022, 3.3]

3.66
machine learning
ML
process of optimizing model parameters through computational techniques, such that the model's behaviour reflects the *data* (3.2) or experience

[SOURCE: ISO/IEC 22989:2022, 3.3.5]

3.67
artificial intelligence system
AI system
engineered system that generates outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives

Note 1 to entry: The engineered system can use various techniques and approaches related to artificial intelligence to develop a model to represent *data* (3.2), knowledge, processes, etc. which can be used to conduct tasks.

Note 2 to entry: AI systems are designed to operate with varying levels of automation.

[SOURCE: ISO/IEC 22989:2022, 3.1.4]

3.68
label
target variable assigned to a sample

[SOURCE: ISO/IEC 22989:2022, 3.2.10]

3.69
training data
data (3.2) used to train a *machine learning* (3.66) model

[SOURCE: ISO/IEC 22989:2022, 3.3.16]

3.70
natural language processing
NLP
<system> *information* (3.1) processing based upon natural language understanding or natural language generation

[SOURCE: ISO/IEC 22989:2022, 3.6.9]

3.71
stakeholder
any individual, group, or *organization* (3.22) that can affect, be affected by, or perceive itself to be affected by a decision or an activity

[SOURCE: ISO/IEC 38500:2015, 2.24]

4 Abbreviated terms:

API	Application Programming Interfaces
ASCII	American Standard Code for Information Interchange
CDO	Chief Data Officer
CEO	Chief Executive Officer
CFO	Chief Financial Officer
CIO	Chief Information Officer
COO	Chief Operating Officer
CSV	Comma-separated Values
EBCDIC	Extended Binary Coded Decimal Interchange Code
JSON	JavaScript Object Notation
RDF	Resource Description Framework (W3C)
RSS	Really Simple Syndication

Annex A
(informative)

Use case template

A.1 Introduction to use case template

A.1.1 General

Use cases were structured using the template described in A.2, however, each case offers a unique perspective on a particular data usage process. For example, use cases can relate to the complexity of the data project, the sensitivity of the data involved or the specific process management mechanisms that were put in place. The focus within these use cases is to examine what measures can be used in data usage projects, particularly in relation to data sharing, to address data privacy, security or integrity issues.

The following are the descriptors for each of the sections presented in the use cases.

A.1.2 Use case name and overview

The use case name and overview provide high level context, with focus more on the data related decisions as they pertain to the data project itself rather than any industry sector specific praxis.

A.1.3 Domain areas

Domain or area to which the use case pertains.

A.1.4 Objectives

Each use case provides context on the objectives of the data project.

A.1.5 Narrative

This section provides a more detailed description of the use case, how it was structured, and the decisions and recommendations that were identified within the use case.

A.1.6 Data lifecycle stages

This section provides context on the data project in terms of the data lifecycle which can be useful for organizations in identifying vulnerabilities within their own data processes.

A.1.7 Figures

This section can assist users of this document in understanding the use case as it is shown diagrammatically. These figures can provide context for the use case, at a high level, as an overview or can reference specific elements within the use case.

A.1.8 Stakeholders and stakeholder considerations

This section can assist users of this document in identifying stakeholders in a data project.

A.1.9 Data characteristics

This section describes the data characteristics within the case study and can include reference to the data type, data representation and data systems.

A.1.10 Key performance indicators

The key performance indicators (KPIs) are described as those elements used for evaluating the performance or outcomes of the data project.

A.1.11 Challenges and issues

The challenges and issues identified within the case study can vary and can include technical data management issues, privacy and security concerns or stakeholder management.

A.1.12 Societal concerns

This section provides commentary on issues that are not commonly considered in scoping a data project as they can relate to issues such as unintended consequences related to how the data outputs are understood, interpreted or applied.

A.1.13 Data security, privacy and trustworthiness

This section provides context as to how security, privacy and trust issues can arise, be identified, and be managed within the data project. The management of data security, privacy and trustworthiness can involve the implementation of levels of control as outlined in Annex C.

A.1.14 Key insights

This section provides context on the insights and learnings from the data project.

A.2 Use case template

The template used for the collation of use cases is provided in Table A.1 and is based on:

— IEC 62559-2

— ISO/IEC TR 24030:2021

— ISO/IEC TR 30176:2021

— ISO/IEC TR 20547-2

Table A.1 — Use case template

Field	Description
1. Use case name and overview	Name of the use case and short overview.
2. Domain areas	Domain or area to which the use case pertains.
3. Objectives	Objectives of the data usage defined by the use case can include the following: — What is to be accomplished. — Intended benefits. — Scope, boundaries and limitations.
4. Narrative	Description, decisions, predictions, recommendations; capabilities and features that are unique to the use case.
5. Data lifecycle stage(s)	Description of the data lifecycle stages related to the use case, decisions made and outcomes.
6. Figures	Diagrams related to the data usage use case. Diagrams can include: — Drawing of a use case. — Data flow diagram of use case. — Sequence diagram of data actions in use case.
7. Stakeholders and stakeholder considerations	Key stakeholders and any considerations, expectations related to the use case.
8. Data characteristics	Describes the characteristics of the data.
9. Key performance indicators	Describes the KPIs for evaluating the performance or outcome of the data usage.
10. Challenges and issues	Descriptions of challenges and issues of the use case.
11. Societal concerns	Describe how societal concerns related to the data use are understood, identified, controlled and mitigated. Entries in this field can include key considerations of Sustainable Development Goals in relation to the use case context; approach taken to digital inclusion considerations; particular attention given to potential harms and sensitivities.
12. Data security, privacy and trustworthiness	Describe the security, privacy and trust aspects of the use case and how they identified, controlled and mitigated.
13. Key insights	Brief description of key lessons and insights to be learnt from this use case. This field is useful both for overall analysis of the use cases and for general application of the use cases.

Annex B
(informative)

Use cases

B.1 Use case 1: Online education analytics

B.1.1 General

This use case describes a commercially developed analytics environment that used proprietary data from an online learning platform to assess online learning activity and outcomes, and the performance of an online education system.

B.1.2 Use case name and overview

This use case provides an overview of a data use project from inception to operation which had several sensitive aspects, including the following:

— students and minors as data subjects;

— data collected direct from data subjects;

— data analysis provided back to data subjects and education institutions;

— automated online interventions based on data analysis.

B.1.3 Domain areas

This use case describes online education reporting and learning analytics (LA).

B.1.4 Objectives

This use case describes the mechanisms for integrating learning analytics, data privacy technologies and ethical practices into a unified operational framework for ethical and privacy-preserving learning analytics. It introduces a new standardized measurement of privacy risk as a key mechanism for operationalising and automating data privacy controls within the traditional data pipeline. It also describes a repeatable framework for conducting ethical learning analytics.

The premise of the project described in this use case is that the existing body of work on the ethics of data management, privacy enhancing technology and ethical analytics can be harnessed to systematize ethical data analytics and reporting. The approach consists in equal parts of education and methodology, secure collaboration and analytics capabilities, and privacy enhancing technology [Figure B.1 a)].

B.1.5 Narrative

This use case describes the use of data by an online education service provider wherein the data source is the service provider's online learning platform (OLP). These data include information about participating students, their online activities and learning outcomes such as:

— engagement with the learning platform and material;

— engagement with other students (in team-based projects);

— engagement with teachers or facilitators;

— assessment details and results.

Many of the programmes involve experiential project-based learning with a real-world client and industry mentor. Feedback and normative grading from these are also included.

The data are put to two uses:

a) Providing experimental data for collaborative learning analytics (LA) research programmes with research collaboration partners. Experimental data include information about participating students, their online activities and learning outcomes and the following:

   1) "time on task" (time spent on each activity);

   2) details of the frequency and sequence of discrete engagements with the learning materials.

   These are only accessible to a small and vetted research team after de-identification (details below).

b) Creating dashboards and reports on student activity and outcomes for internal, external client and public consumption.

   This is mostly summarized information, however, some drill-down data is available to customers. This includes:

   1) Internal performance reporting: Platform and program usage, student engagement, program-related ratings, billing-related data.

   2) Internal product reporting: Feature usage and engagement.

   3) Customer reporting (mainly teaching institutions): Usage and engagement for their institution, student performance.

   4) Whitepapers for general publication: Studies of student engagement, completion rates and outcomes.

The key intended outcomes include the following:

— improved understanding of behaviours conducive to experiential learning and ability to use online learning data as a proxy;

— understand engagement with the learning platform and student outcomes to improve platform features and learning programme design;

— develop public domain whitepapers on student outcomes and other benefits of the platform;

— standard business metrics reporting.

B.1.6 Data lifecycle stages

B.1.6.1 General

The data lifecycle stages for this use case are described against a four-tiered data management structure [Figure B.1 b)].

B.1.6.2 Extraction and de-identification

This stage maps to "2. Extraction and De-identification" in Figure B.1 b).

See also the data management quadrant in Figure B.1 a).

Data were not collected from the end user for this use case. Semi-structured data were extracted from the source learning platform by an internal API. Basic de-identification was performed as part of this process.

Direct identifiers, such as student name and ID, phone numbers, etc. were replaced by non-reversible unique identifiers (UIDs) or removed altogether. The resulting de-identified data were stored in a data lake.

a) Control environment level: High control.

   Data extraction for analytics was done in accordance with the organization's data privacy and data use policies, and approval was provided via client agreements. The purpose of collection and use was clearly stated.

b) Principles:

   1) Separation of uses (a key principle). Source system data was not available in its "raw" form for anything other than the learning platform's primary use. That is, the development, management and support of a learning programme and enrolled participants. Roles authorized to access the source system data were limited to:

      i) authorized learning platform users (within the scope of their role);

      ii) authorized customer support personnel;

      iii) devOps technical support personnel (when required to support a customer or the platform).

      Other uses of the data were designated "downstream". Downstream users did not have access to the learning platform database or "raw" data.

      Analysts working on dashboards and reporting, or learning analytics research were able to access extracted and de-identified copies of the data from the data lake. Further user segmentation existed based on the user's role.

   2) Privacy by design (PbD). The separation of uses and de-identification of data for downstream use was key to PbD. Automated de-identification begins to build privacy principles into the data pipeline infrastructure.

B.1.6.3 Privacy risk assessment

This stage maps to "3. Privacy risk assessment" in the data management structure (Figure B.1 b)).

See also quadrants 2 and 3 in Figure B.1 a).

The data went through two more preparation steps before they were used:

— Privacy risk assessment. This was an automated procedure in which data were passed through a privacy risk assessment tool which derived a privacy risk evaluation (PRE) metric. PRE is a statistical measure of the risk that an individual can be re-identified from the data.

  The PRE calculates a numerical measure of the risk of re-identification of a dataset which has had preliminary de-identification already applied (removal of names and other direct identifiers). The PRE tool uses a combination of statistical and Artificial Intelligence techniques to arrive at this measure[37].

— Metadata enhancement. The source extract process populated basic metadata values such as data sources, and date and time data. More metadata were added to provide a minimum level of information about the collected data (including source details and timing, completeness, treatment or transformations applied, fitness), depending on context.

a) Control environment level: High control.

   Authority to store was controlled by the organization's data management policies and procedure which dictated the PRE levels required. Metadata included PRE and data provenance and completeness information.

b) Decisions:

1) PRE threshold failure. An automated threshold was set at this point to highlight data that had a higher-than-desired PRE (i.e. above the set re-identifiability threshold). There needed to be a decision at this point to:

i) apply additional privacy enhancing techniques to the data;

ii) remove the data; or

iii) store the data without change.

2) Data characteristics. Analysts were able to decide how to treat data to the extent that there was the possibility of the analysis results being skewed or being considered not-representative. There needed to be a decision at this point to:

i) withhold access until problems were able to be rectified;

ii) remove the data; or

iii) store the data without change (but include metadata information about fitness).

c) Principles:

1) Privacy by design (PbD) and default. Inclusion of PRE measures with all data was part of the automated data storage process. Application of additional privacy enhancing techniques as necessary was either manual or automated. All uses of the data were informed by its PRE and metadata.

2) Ethical data usage. Metadata and PRE helped inform analysts on the ethical implications of their data projects. For example, the PRE assisted with decisions about the release of data to other parties [Figure B.1 d)]. Metadata assisted in appropriate selection and use of data for analytics and reporting.

B.1.6.4 Analysis and Reporting

This stage maps to "4. Analysis and Reporting" in Figure B.1 b).

The approach to learning analytics and general reporting was similar, governed by the operational components depicted in Figure B.1 a).

a) Control environment level: High control.

Explicit authority to use and release outputs was required. PRE measures were consulted for decision-making on data sharing. Expertise and experience of analysis and data scientists was ensured.

b) Decisions:

1) Ethical purpose. This primarily pertained to predictive analytics and profiling or categorization projects. It required the project to be justified based on its beneficial purpose and theoretical underpinnings. This was supported with an experimental design framework to prompt analysts to think about their decision making around data selection, treatments and the application of results, and an open peer review process.

2) Reports and dashboards. This included the PRE metric for the underlying data to inform the recipient on their decisions to share reports or to copy or reuse data.

c) Principles:

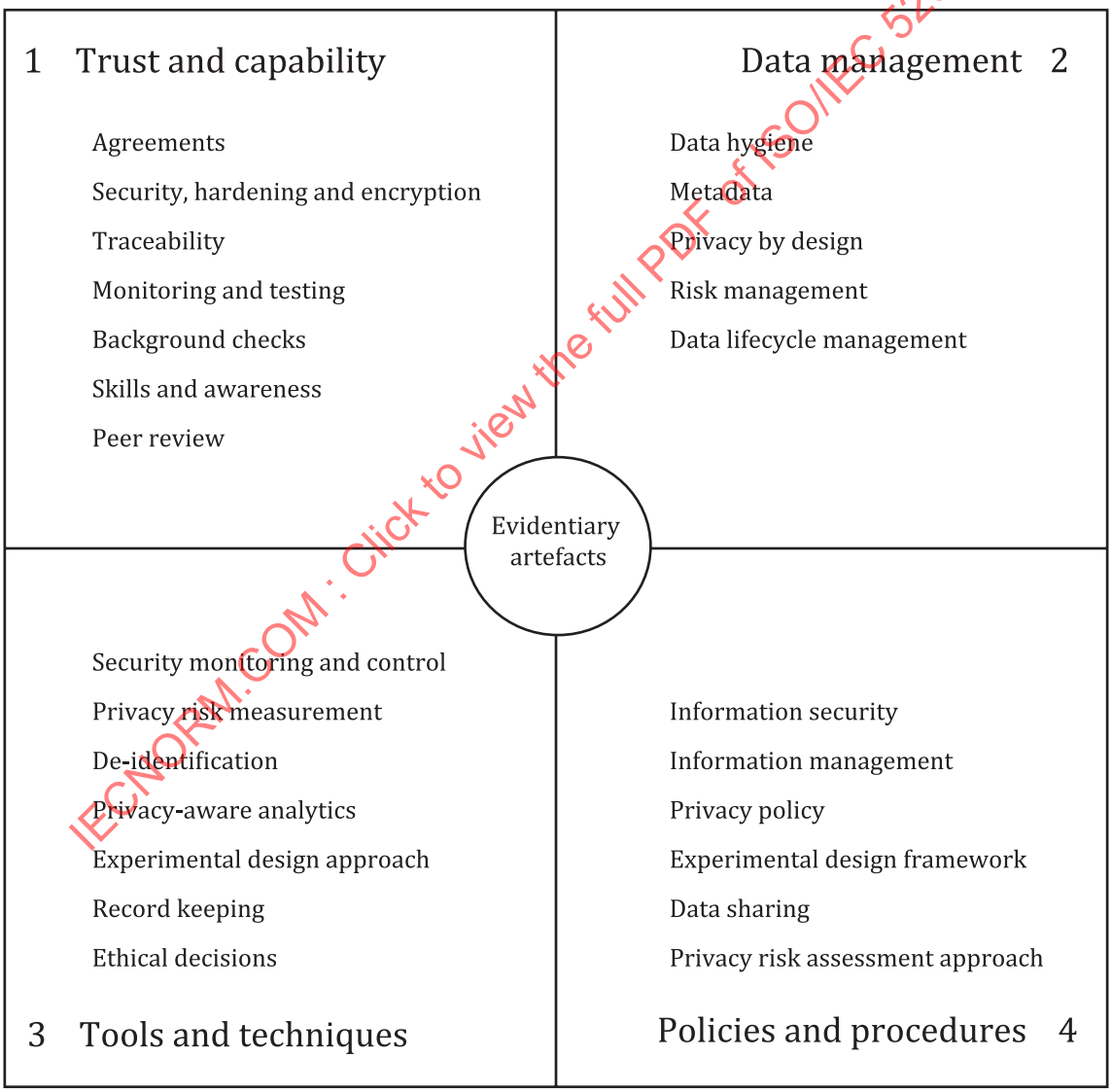
1) Privacy by design. Continued reference to and re-use of PRE brought privacy considerations to the fore in the preparation and distribution of reports. Report recipients were reminded of data privacy with the inclusion of PRE measures in dashboards and reports.

Collaborative projects were required to work inside the secure collaboration environment to protect data and intellectual property [Figure B.1 d)].

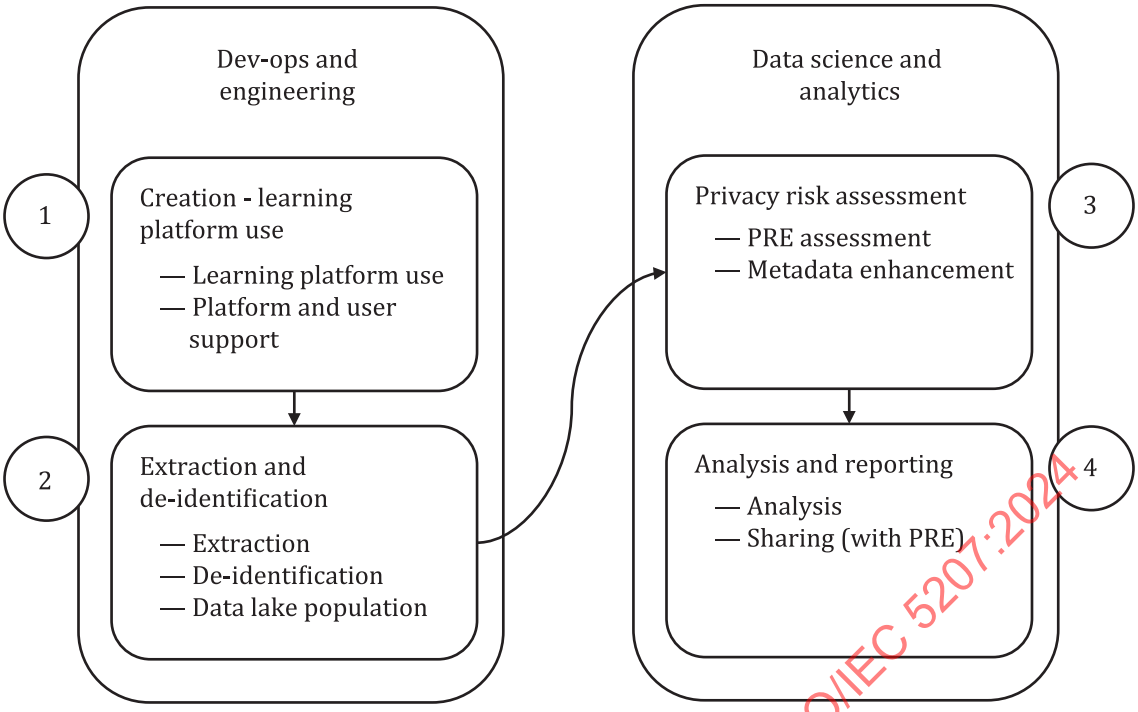
2) Ethical decision making. A supporting experimental design framework ensured consideration and documentation of the impacts of algorithm and data selection as well as treatments and omissions.

NOTE Marshall et al[37] pp 748-751 contains a detailed example of the methodological components of an ethical decision making framework including details of an experimental design framework.

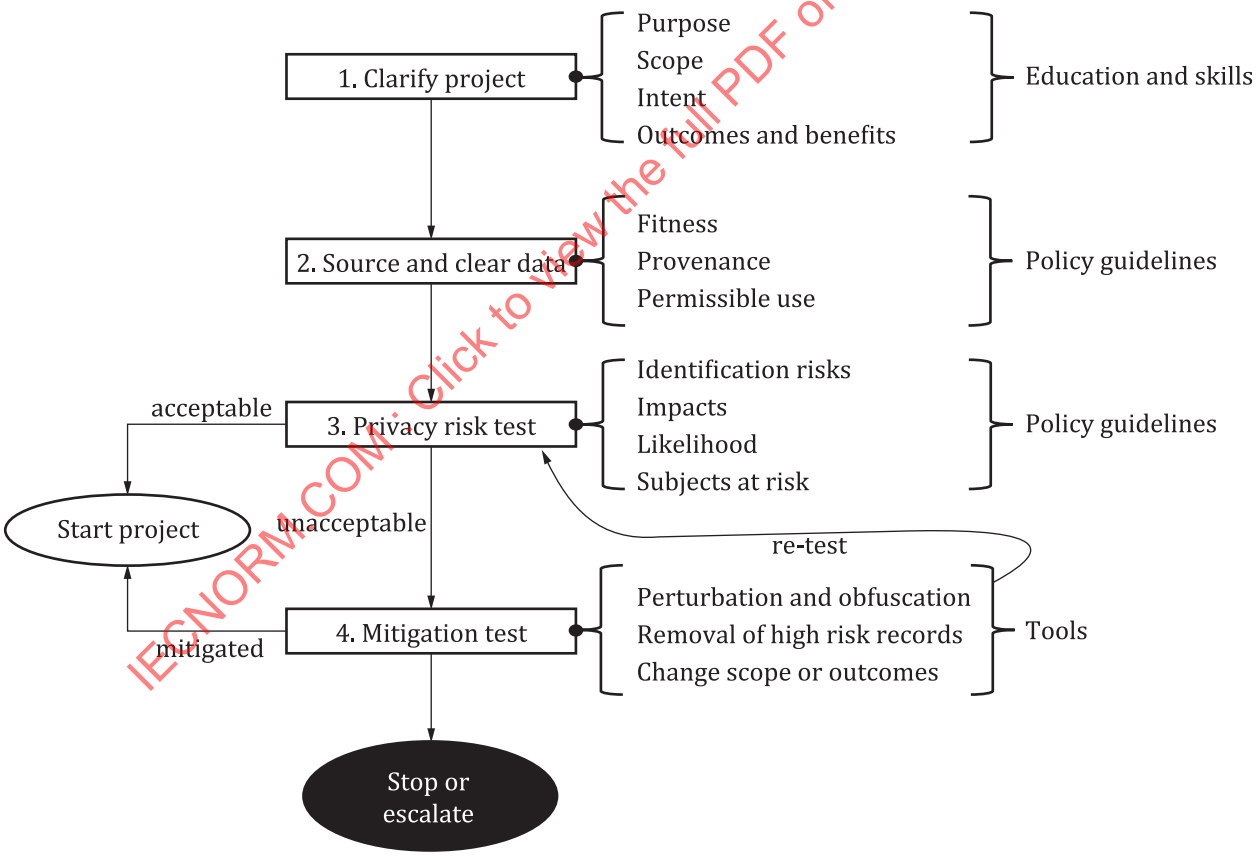
B.1.7 Figures



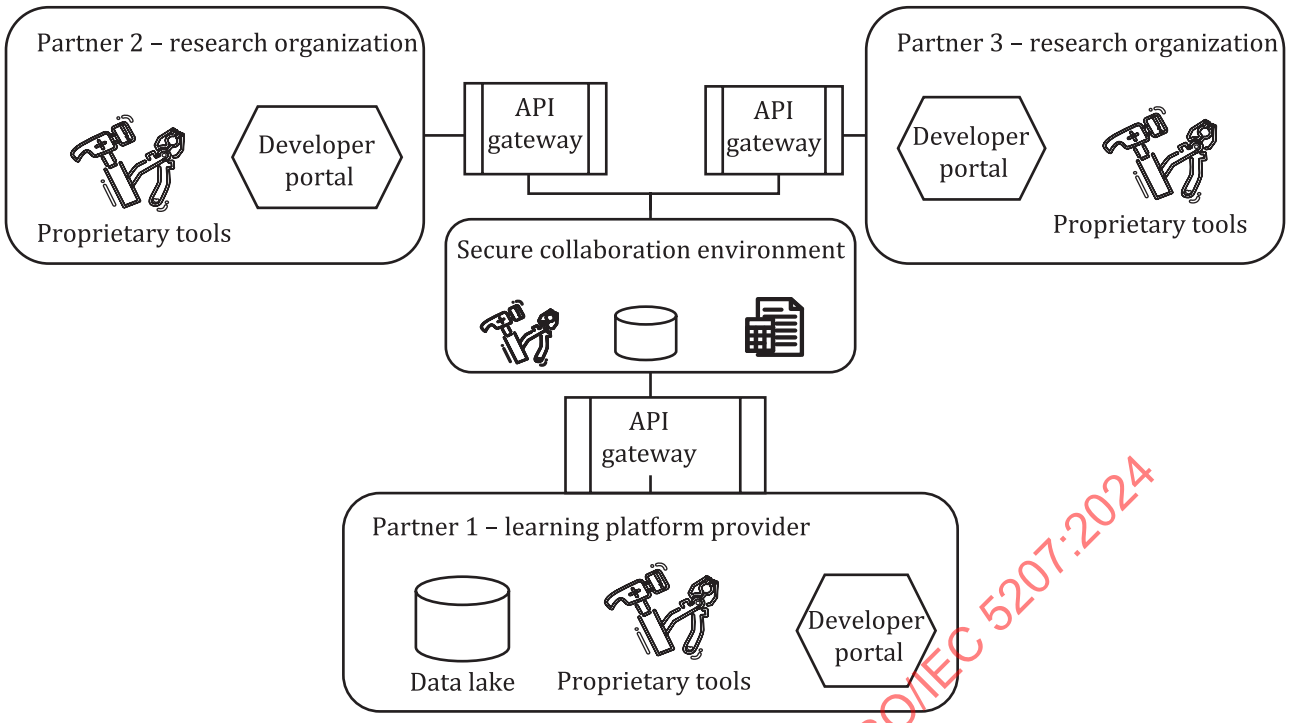
a) Key components of an operational framework for ethical data analytics



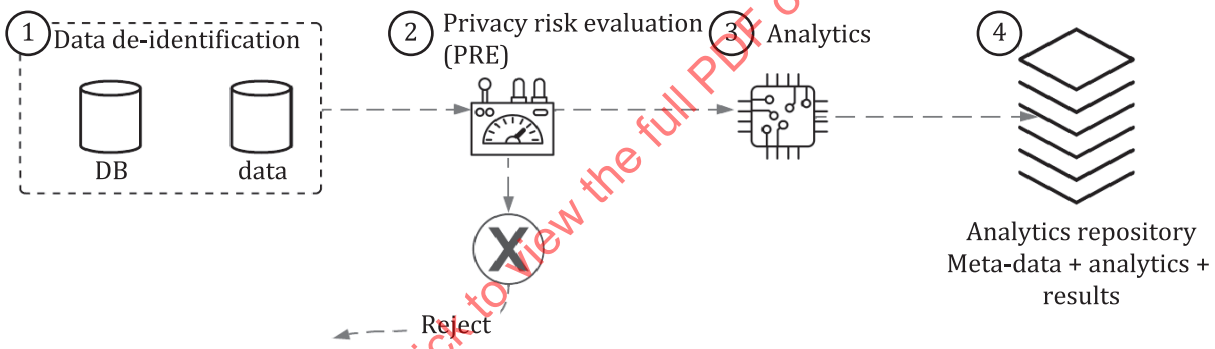
b) Four-tiered data management structure



c) Privacy risk assessment decision flow, part of a wider methodological framework



d) Secure collaboration, experiment working scenario



e) Example experimental pipeline

Figure B.1 — Use Case 1

B.1.8 Stakeholders and stakeholder considerations

The key stakeholders included the following:

— Corporate learning platform and services provider. An online experiential learning platform and services provider, providing skills development programmes to adults and university students. Over fifty percent of Australian universities contemporaneously engaged the services of this organization, with many thousands of students enrolled.

— Corporate clients. Higher education institutions, mainly universities, which used the learning platform and services to provide skills development programmes to their student cohorts.

— Students. Higher education students who engaged with the learning platform as part of in-curriculum and extra-curricular skills development programmes at their university.

— Research collaborators. Industry partners and research institutions who worked with the corporate learning platform provider to conduct research into learning analytics and privacy enhancing technologies for the online learning industry.

— Government sponsors. Financial support was provided for a major collaborative research project by the Australian Department of Industry, Science and Resources.

Stakeholder considerations included the following:

— Student data privacy and protection. This was a key consideration for all stakeholders, and in particular the learning platform service provider which was the main source of data. The analytics conducted with student learning data offered significant potential benefits to the business and to future students via targeted interventions and other platform improvements. Both legally and ethically, these benefits cannot be achieved at the expense of a breach of student privacy.

— Data and intellectual property protection. Data sharing between stakeholders during research collaborations and report distribution was a key part of the programme. Management of data and insights, including physical and logical separation between collaborators as well as client organizations was a key consideration.

— Avoidance of adverse impacts. Validation and monitoring of insights and interventions based on learning analytics was used in order to avoid adverse or negative impacts. Risk management plans to respond to unintended adverse impacts was key to this consideration.

B.1.9 Data characteristics

Data characteristics for this use case included the following:

a) Student data. Un-summarised, time series data describing student activities and engagement with the material on the learning platform included:

   1) Enrolment information included basic student identification, university or other learning institution, and start and end dates.

   2) Reading and engaging with instructions and teaching content.

   3) Progress against tasks and assessments.

   4) Engagement and discussion with other students, mentors, and student supervisors.

   5) Student submissions included written material, video submissions and online assessments.

   6) Normative and informative feedback from mentors and supervisors in the form of grades or comments.

   7) Student ratings of the learning experience, platform, and the performance of mentors and other coordinators.

b) Course data. Course design information included structure and learning outcomes goals.

c) Metadata. At a minimum the following:

   1) Contextual information about the data including date of collection, some programme and institution information, relevant fitness information (such as completeness).

   2) Privacy risk evaluation metric associated with the data.

As described in this use case, data were subjected to rigorous privacy enhancing techniques from the start and individual student identifiers were removed or replaced. The nature of much of the analytics required that pseudo-identifiers were used so that it was possible to analyse discrete student journeys.

B.1.10 Key performance indicators

Key performance indicators for this use case included the following:

— Standardization of data privacy management for all application data, without significant impact on productivity or ethical use.

— Systematic application of a high standard of care and responsibility in data analytics and reporting activities that were verified via documented decision making at key ethical decision points (project justification, data selection, treatment, verification, etc.) and peer review.

— Built-in mechanisms to support data privacy by design and default with the automation of key initial de-identification and separation of duties.

B.1.11 Challenges and issues

Challenges and issues for this use case included the following:

— Up-front establishment of infrastructure and procedures. There was some front-loading of project costs to establish the security and privacy by design infrastructure. Although there was a good business case for doing so, with the downstream, efficiencies from a semi-automated process, it was a difficult investment case to make, especially as this was not standard practice at the time.

— Establishment of a harmonized PRE. There was continued active research to develop the underlying techniques to measure privacy risk. With a number of approaches used, each having strengths in different areas, the harmonization of these approaches into a single mechanism for measurement was challenging.

— Development of PRE thresholds. Trial and error and a high volume of testing were required to assess the PRE thresholds for different levels of data access. Every organization had different circumstances, relationships with their customers, and risk appetite, and needed to go through a similar process. Longer term, industry standardization made this process more efficient and facilitated better data sharing agreements and procedures.

B.1.12 Societal concerns

Societal concerns for this use case included the following:

— Representational qualities of the data. Most programmes which provided data for this use case were extra-curricular skills development programmes targeting undergraduate and graduate overseas students. Participation was voluntary and therefore the student cohort was self-selected. It was therefore important to understand how the resulting analytics can and cannot be used (how generalizations can be managed for example). It was also important to make use of a detailed experimental design and an experimental pipeline [Figure B.1 e)] which preserved data and experiments for future replication and revalidation.

— Learning outcomes. Many of the learning programmes were project based experiential learning programmes where student teams conducted short real-world projects for an industry organization. Outcomes were often based on the success of the project and feedback of the business mentor or project sponsor. It was important to ensure that this was not used as a proxy of student learning outcome without more direct data (such as a learning assessment component).

— Student agency. An ethical consideration when using analytics to optimize the online learning process was to balance improved student outcomes with student agency. Student agency, as defined by the OECD is "…the capacity to set a goal, reflect and act responsibly to effect change", making students more able to "learn how to learn"[38]. This was a consideration when experimenting with automated interventions based on analytics insights. It was important to ensure that all such research was supported by learning theory and student's agency was not removed. In most cases, the team leaned towards using insights to alert student supervisors or coordinators to suggest an intervention in preference to automating the intervention directly.

— Data privacy and security. As discussed in B.1.14, data privacy was particularly important when analytics created new insights about individual student performance. The default position was not to surface insights about individual students.

B.1.13 Data security, privacy and trustworthiness

Data privacy, security and trustworthiness were key components of the project, its goal being to develop actionable and practical privacy enhancing learning analytics for the education industry.

Results for this use case included the following:

a) Privacy by design. As illustrated in Figure B.1 b), data privacy was built into the fabric of the end-to-end data pipeline to enable privacy by design and default:

   1) Separation of duties. A key part of this was to support the separation of duties so that only those supporting the platform and its primary uses had access to the raw, identifiable, data. Analysts and other down-stream users of the data only saw de-identified information.

   2) Pseudonymization. Removal of direct and quasi-identifiers at source, and replacement with non-reversible identifiers where necessary.

   3) Privacy risk evaluation (PRE). A mathematical measure of the risk of re-identification was stored with the data at all times to inform users and recipients of the data on their treatment and sharing. PRE measured the residual risk of re-identification once privacy enhancing techniques had been applied. Figure B.1 e) shows a simplified example of the automated use of PRE in the experimental pipeline established for collaborative research. Because this was a collaboration with external research partners, the data were not made available to researchers if they failed to meet a pre-set threshold at step 3 of Figure B.1 e). Further privacy risk reduction treatments had to be applied before the data were available to the project.

   4) Privacy awareness and literacy. Corporate data management procedures and education assisted analyst and report recipients to make sound decisions about sharing and use of the data based on the PRE. Report recipients, including professionals, were visually reminded of the people behind the data and residual privacy risk by attaching the PRE to reports and dashboards.

b) Perimeter security for collaboration and data sharing. A cloud network security environment was used for research collaborations and also to provide detailed student information and reports securely to institutional customers. Collaboration and terms of use agreements were required for project participants and customers.

   Using a commercial suite of tools, a multi-party secure collaboration environment was established where all collaborative research was conducted. Access was controlled with 2-factor authentication using a hardware-based solution and regular user reviews. Movement of data in and out of controlled areas was logged. Secure API connections enabled collaboration between participant organizations across sovereign areas, as well as common areas which were accessible by all participants. All movement of data in and out of the environment, and between secure areas, was logged, enabling a “trust but verify” approach to bolster contractual collaboration and data sharing agreements.

   The working scenario in Figure B.1 d) gives an illustration of how a collaborative group comprising three different organizations engaged in a collaborative research project to develop privacy enhancing learning analytics tools, which is documented in Marshall, et al. (2022)[37].

   Background police checks were also mandatory for all employees and contractors with access to the data or application back end, without which there was a risk of other security measures being invalidated.

c) Trustworthiness. Secure environments and a privacy by design approach both served to contribute to trust (trust but verify). These supported basic trust structures, such as legally binding collaboration agreements and terms of use.

The addition of PRE metrics to all externally reported data was a measure to deepen a sense of trustworthiness with the customer base.

B.1.14 Key insights

This use case was an experiment in merging existing tools and methodologies to “bake in” the ethical treatment of data across the entire data lifecycle. Sufficient progress has been made in key areas of data

privacy, security, metadata management, experimental design and ethical considerations around the appropriate use of data that a more ethical, wholistic approach to ethical data management and use is possible today. This will improve over time as more tools become available.

To a great extent, the process can be systematized to remove, where appropriate, the friction that privacy and ethics processes can introduce and to give analysts and users more freedom to innovate. This does not simply mean the mechanization or automation of processes, but having associated methodologies and decision-making frameworks and broader education around ethical data use.

This approach can be particularly applicable to the small to medium enterprise and start-up business communities, which have more limited focus, budget and expertise for managing their data responsibly. Shared tool sets and standardization on PRE thresholds can pave the way for more effective data management for these organizations.

B.2 Use case 2: Government led COVID-19 case data publication

B.2.1 General

Use case 2 was a government data publication programme that included sensitive data, de-identification of data and statistical analysis to provide a level of assurance around the risk of re-identification. Data was published daily and made publicly available in a highly sensitive environment commencing with the first major wave of COVID-19 infections.

B.2.2 Overview

Use case name: Release of COVID-19 confirmed cases as open data in New South Wales (NSW), Australia

This use case provides an overview of a data project which published current de-identified sensitive medical data (COVID-19 cases) taken from government health services information. A mathematical treatment was used to analyse the data prior to publication to determine whether re-identification of the data subjects was possible. This mathematical treatment was referred to as a personal identification factor (PIF).

B.2.3 Domain areas

Government data and the public domain: Release of COVID-19 confirmed cases information as open data daily by postcode in NSW, Australia from March 2020 to August 2022.

B.2.4 Objectives

The objective was to increase situational awareness of the developing COVID-19 situation in NSW Australia by releasing data at a fine-grained spatial (postcode) and temporal (daily) level of confirmed cases, likely source of infection, age of infected individuals, as well as records of COVID-19 testing.

B.2.5 Narrative

In March 2020, the NSW government committed to release information about the developing number of confirmed COVID-19 cases on a daily basis at postcode level. Issues of the level of personal information and the sensitivity of the data were of foremost concern. This was balanced with the strong desire of the public to be informed about the developing COVID-19 situation. A complete set of possible fields for release was collated from NSW Health sources, and then tested for the total amount of information that would be revealed about individuals if released (and the individual was identified).

A series of consultations were undertaken regarding the balance of data being released “in the public interest” versus data which were merely “of interest to the public”. The risks associated with re-identification of individuals was also considered and how much information could be associated with an individual who was identified.

A personal information factor (PIF) tool had been developed and tested some time earlier. The PIF tool was used to develop an upper limit measure of the worst-case (greatest amount of) information that would be

released if an individual were identified. This tool and measurement process was used to design in additional protections (disconnecting features, aggregation, obfuscation) for the data before they were released as open data. The data in the reduced feature tables were analysed each day to ensure the PIF was reduced to an agreed level before release. The data set was assumed to be in the form of rows (unique individual) and columns (features related to that individual). The data released were also used to create spatial maps for those who did not want to access the data directly.

For more information, see the NSW Government case study "Case Study: Personal Information Factor (PIF) Tool"[39].

For a description of the PIF tool, see the CSIRO and Data61 web page[40].

B.2.6 Data lifecycle stages

Data traverses through the stages of collect, transmit, store, analyse in a high control environment. The data products created from analysis are released into a high control environment, with access limited to authorized individuals for respective environments, or a no control environment released to the wider public as open data. Figure B.2 a) shows the data lifecycle stages in relation to various levels of control in the environment. Within this control environment, two data product sets were created:

— High control environment: Raw data with personal information and sensitive location information only accessed by data custodians.

— No control environment: Raw data reduced levels of personal information and spatial information at postcode level.

A very high control environment was not required as data collection and use did not require a special legal instrument. The path of these two data products through the data lifecycle stages is shown in Figure B.2 b).

Control levels were calculated based on (proven) capability, (assessable) governance and (verifiable) purpose. Doing so required an objective, repeatable, standardized assessment of:

— capability;

— governance;

— purpose;

— data quality and provenance;

— sensitivity of data;

— degree of personal information contained in datasets.

A no control environment requires no assessments nor any restriction on people accessing or utilising data. A high control environment involves skilled people working in a strong governance environment with clearly authorized purpose. It is also important to note that capability included skill in all stages of the data lifecycle: data analysis, data provenance, governance and security.

Figure B.2 a) illustrates how different levels of control in the environment will have different requirements in relation to the authority to collect, use or share data, metadata about data quality and provenance, and expertise relevant to the data, project and data lifecycle stage.

A very high control environment has:

— high quality data and metadata;

— specific purpose and authority to access and use data;

— expert users experienced with the quality of the data provided and with associated metadata;

— expert analytical capability and domain expertise;

— strong governance and security at each stage of the lifecycle;

— specific restrictions on release of data and insights, or secondary use of data and insights;

— people that have met general expertise requirements as well as project specific requirements and agree to be bound by limitations on data access and use.

Data in this environment is very high sensitivity with a very high PIF. This level of control is suitable for:

— data which can only be accessed under an external instrument such as a Public Interest Direction (PID);

— data which is reasonably personally identifiable;

— data which contains sensitive subject matter;

— data which has a well quantified quality (need not be high quality).

A high control environment has:

— explicit purpose and authority to access and use data (although it might not have project specific requirements);

— expert users that are experienced with data of the quality provided, and with associated metadata about data quality and provenance;

— very skilled analytical capability and domain expertise;

— strong governance and security at each stage of the lifecycle;

— explicit restrictions on release of data and insights, or secondary use of data and insights;

— people with access that have met general expertise requirements and have agreed to be bound by limitations on data access and use.

In such an environment, there can be broad authority to collect, store, and use data. Data in this environment is moderately sensitive with a moderate PIF. This level of control is suitable for:

— data which is not reasonably personally identifiable;

— data which contains sensitive subject matter;

— data which has a well quantified quality.

A moderate control environment has:

— general purpose and authority to access and use data (such as an authorizing regulatory framework);

— experienced users dealing with data of the quality provided and with associated metadata about data quality and provenance;

— skilled analytical capability and domain expertise;

— strong governance and security at each stage of the lifecycle;

— general restrictions on the release of data and insights, or secondary use of data and insights;

— people with access that have met general requirements and have agreed to general conditions on data access and use.

In such an environment, there would also be a broad authority to collect, store and use data. Data in this environment is moderately sensitive with a moderate PIF. This level of control is suitable for:

— data which are not reasonably personally identifiable;

— data which contain some sensitive subject matter;

— data which areof sufficiently high quality for the intended use.

A low control environment can have:

— no explicit authority to collect and use data, but no known restrictions to use data;

— users with some experience dealing with data of the quality provided;

— users with some analytical capability and domain expertise;

— appropriate governance and security at each stage of the lifecycle.

A low control environment can have assumed authority to collect, store, and use data and can have metadata on data provenance and quality. In this use case, such an environment did not have restrictions on release of data and insights, or secondary use of data and insights. In this environment, data has a low PIF. This level of control is suitable for:

— data which are not reasonably personally identifiable;

— data does not contain sensitive subject matter;

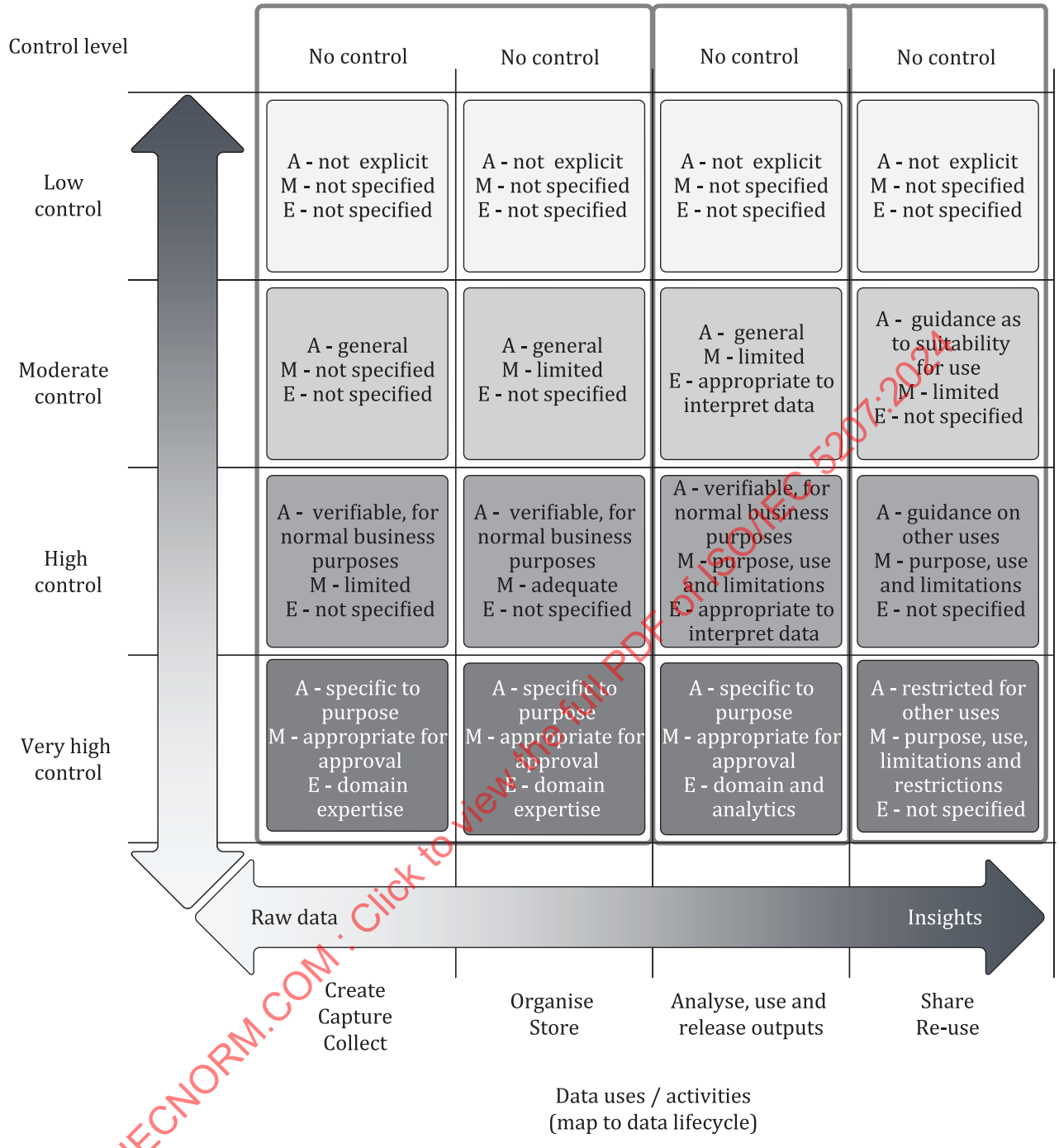
— data which are of sufficiently high quality for general use.

A no control environment can have no controls in place. It is suitable for:

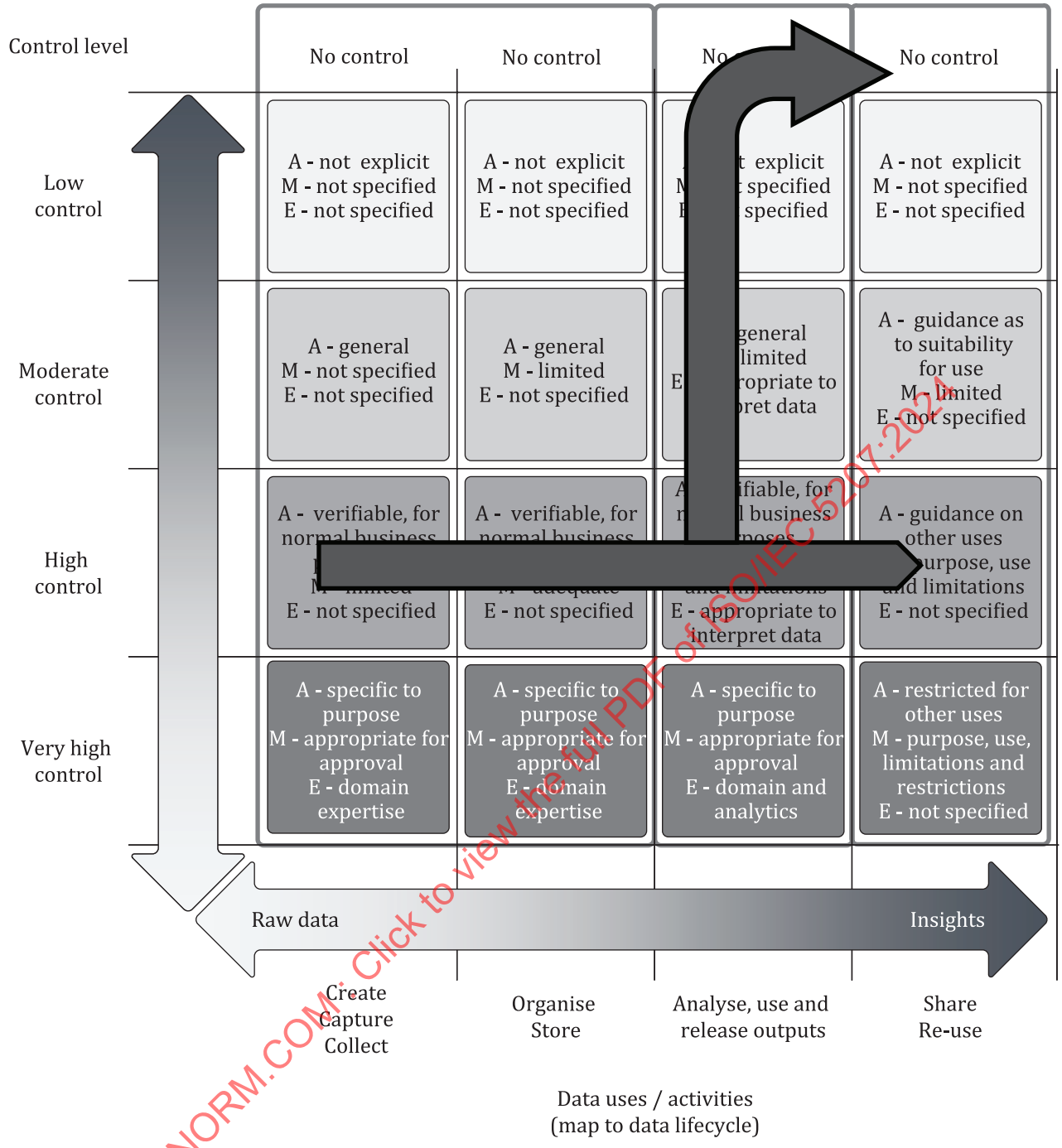
— data which have been approved for release as open data;

— data which are of sufficiently high quality for general use.

B.2.7 Figures



a) Layers of control for data access and use, as well as data product on-sharing



b) COVID-19 confirmed cases data and data product lifecycle with arrows overlaying table depicted in subfigure a) to show path of these two data products through the data lifecycle stages

Key

A authority
M metadata
E expertise

Figure B.2 — Use case 2

B.2.8 Stakeholders and stakeholder considerations

The key stakeholders include the following:

— data custodians at Ministry of Health, Premier and Cabinet and Customer Service;

— domain experts from contributing agencies;

— data analysis team at NSW Data Analytics Centre (DAC);

— NSW Chief Data Scientist;

— cyber security and governance teams at NSW DAC;

— NSW Privacy Commissioner;

— NSW Ombudsman.

The stakeholder considerations related almost exclusively to the reliability of the data and risk of re-identification (privacy and security).

B.2.9 Data characteristics

The data characteristics included the following:

— unit record level, de-identified data containing sensitive health information and sensitive personal information;

— data products at aggregate (insights level).

B.2.10 Key performance indicators

Data of this type (unit record health data released at fine grained temporal and spatial levels) had never previously been released to the public. KPI's were based on NSW customer feedback, complaints and input from the NSW Privacy Commissioner and the NSW Ombudsman.

B.2.11 Challenges and issues

The data set was updated daily for approximately 2.5 years. Based on NSW customer feedback, the data were used for a range of local reporting applications. Very few complaints were received. The complaints that were received were addressed by developing guidance for use of the data released to the public.

B.2.12 Societal concerns

The data used to create the COVID-19 confirmed cases data asset are protected by NSW state legislation and government policies. The declaration of a health emergency enabled unusual levels of data sharing within NSW government and with the public. Normal measures were applied to data use that ensured protection of privacy and that data were kept secure, including a comprehensive framework of security controls and standards.

B.2.13 Data security, privacy and trustworthiness

The data used to create the COVID confirmed cases data asset are protected by NSW State legislation and government policies. The declaration of a Health Emergency enabled unusual levels of data sharing within NSW Government and with the public. Normal measures were applied to data use that ensured the protection of privacy and that data was kept secure, including a comprehensive framework of security controls and standards.

B.2.14 Key insights

Given the nature and granularity of the dataset created for situational awareness for COVID-19 case development in NSW, each sensitivity identified was "high".

Aspects of sensitivity of use of data included:

— sensitive subjects recorded in data (health data);

— fine grained temporal reporting (daily);

— fine grained spatial reporting (postcode).

When data were released each day, the entire data asset (back to March 2020) was reviewed for its PIF, ensuring that the concerns of re-identification addressed every element of data which had been released back to first release.

B.3 Use case 3: Cross government agency data sharing to support human services system reform

B.3.1 General

This use case concerns a data sharing project involving numerous government agencies at state and Commonwealth (federal) level in Australia to identify areas for improvement in the case management of children in out of-home-care (OOHC).

B.3.2 Use case name and overview

Use case name: Human services data set — Underpinning OOHC reform in NSW, Australia

This use case outlines the steps taken to provide greater security, safety and privacy protection in the development of a shared database which drew on source data within a number of government agency data sets.

B.3.3 Domain areas

The domain areas included in this use case came from 11 government agencies, including Family and Community Services, Child Services, Health Services, and Justice. However, this use case can provide guidance to any data sharing project, particularly ones involving services for children and young people, or any other projects with sensitive personal information or personally identifiable information.

B.3.4 Objectives

Key to the system reform is the underpinning data set, now known as the Human Services Data Set, or HSDS. It brings together data collected by individual NSW government agencies to take a unique and powerful view of service usage and effectiveness to improve outcomes. It links administrative data across Education, Health, Justice, Industry, Transport, and other agencies. The data set consists of linked longitudinal data "journeys" of individual children and their families. Concerns about increased numbers of data sets leading to greater risk of creating Personally Identifiable Information (PII) motivated the design of the shared database described in this use case. See the concept diagram shown in Figure B.3 a) for an illustration.

The separate data were de-identified and linked by a specialist data linkage centre, meaning all records were anonymous.

The data were, nonetheless, personal in nature and sensitive from many perspectives.

B.3.5 Narrative

B.3.5.1 Overview

In 2015, an independent review of OOHC examined the state of the system in NSW. The review found the system to be ineffective and unsustainable, failing to improve long-term outcomes for children or to arrest the devastating cycles of intergenerational abuse and neglect. Outcomes were particularly poor for Aboriginal children, young people, and families.

The review outlined a vision for whole-of-system reform to deliver improved outcomes for vulnerable children, young people, and families. This recommendation was approved by NSW cabinet in August 2016 and led to the establishment of the Their Futures Matter (TFM) programme.

In the first three years of reform, NSW progressed each of the following recommendations of the 2015 review:

— developed and implemented an investment and commissioning approach to direct funding, effort, and resources to those with the greatest needs;

— established a cross-agency dataset and investment model which provided a full picture of total system expenditure for vulnerable children and families;

— developed a framework to monitor and evaluate outcomes for vulnerable children and families across government;

— expanded investment in intensive family preservation and restoration services using the evidence-based models: Multisystemic Therapy for Child Abuse and Neglect (MST-CAN) and Functional Family Therapy – Child Welfare (FFT-CW);

— developed a trauma treatment service for children in OOHC, and new investment in sustaining OOHC placements to improve stability for children;

— introduced personalized support packages for identified cohorts of vulnerable children and their families;

— improved planning and support for young people transitioning to adulthood from the OOHC system;

— increased the evidence base for interventions that work for Aboriginal children, young people, families, and communities;

— redesigned the intake, assessment, and referral 'access system'.

Key to the success of the investment approach was the underpinning data set, known as the HSDS. It brought together data collected by individual government agencies to take a unique and powerful view of service usage and effectiveness to improve outcomes.

The separate data were de-identified and linked by a specialist data linkage centre, meaning all records were anonymous.

B.3.5.2 About the data

The data set was unprecedented in scale in NSW, bringing together 27 years of data, over seven million records from over 60 frontline data sets in 11 government agencies.

The service streams, outcomes and life events included in the data model were as follows:

a) Child protection:

   1) Concern reports.

   2) Risk of Significant Harm (ROSH) reports.

   3) Safety Assessment, Risk Assessment and Risk Reassessment (SARA).

   4) OOHC episodes (own and next generation).

   5) Number of placements in OOHC.

   6) OOHC placement type.

   7) Primary issue given as reason for concern report and SARA.

8) Restoration.

b) Housing:

1) Social housing tenancies.

2) Private rental assistance.

3) Homelessness services.

c) Justice:

1) Custody.

2) Community supervision.

3) Court finalizations.

4) Cautions.

5) Youth conferences.

6) Legal aid.

d) Health:

1) Public hospital admissions.

2) Private hospital admissions.

3) Emergency department presentations.

4) Ambulance patient contact events.

5) Childbirth.

6) Opiate treatment programme.

e) Education:

1) National Assessment Program — Literacy and Numeracy (NAPLAN) year 3 results.

2) NAPLAN year 7 results.

3) HSC completion.

4) Unexpected government school moves.

5) Resource Allocation Model (RAM) equity loadings.

f) Mental health:

1) Hospital admission for mental health.

2) NSW ambulatory mental health.

g) Alcohol and other drugs (AOD):

1) Hospital admission for AOD.

2) Proven AOD offences.

h) Parental risk indicators:

1) Parent in custody.

2) Parent interaction with justice.

3) Proven AOD related offence or AOD hospital admission.

4) Proven domestic violence related offence or victim of domestic violence.

5) Treatment for mental health in NSW hospital or ambulatory services.

i) Commonwealth services:

1) Welfare.

2) Medical Benefits Scheme (MBS).

3) Pharmaceutical Benefits Scheme (PBS).

This HSDS, with integrated, de-identified data from across six NSW clusters (Health, Education, Justice, Family and Community Services (FACS), Industry and Treasury) will enable detailed planning into what investment and resources are needed in the future, and where effort should be prioritized. Creation of the data set required extensive coordination between policy, legal and data stakeholders in human services agencies and input from the NSW DAC. The data set provides a rich linked history of each child and related persons.

B.3.5.3 Creating the HSDS

The individual agency datasets used to create the HSDS were linked using a master linkage key approach by the Centre for Health Records Linkage (CheReL).

Record linkage brought together information that related to the same individual, family, place, or event from different data sources. In this way, it was possible to construct chronological sequences of health events for individuals. Combined, these individual 'stories' created a larger story about the health of people in NSW and the Australian Capital Territory (ACT).

Once the relevant government agencies and a human research ethics committee approved the research project, the data were then able to be securely linked by the CheReL and made available to the researcher.

The technique used to create the HSDS was a Master Linkage Key system of continuously updated links within and between core de-identified health-related datasets. Both person and family-based extractions were available, as were tailored linkage services using additional datasets.

Quality assurance checks were regularly carried out on the Master Linkage Key. For more information on quality checks and procedures, see the TFM Quality Assurance web pages.

B.3.5.4 Governance

The data used to create the HSDS are protected by laws and other measures that guard privacy and keep the data secure, including a comprehensive framework of security controls and standards.

B.3.5.5 Privacy

The NSW HSDS was created by de-identifying and combining data collected through the administration of different NSW Government services and some Commonwealth Government supports (i.e. welfare and medical benefits). Information like names, dates of birth and addresses were removed to ensure that the data did not identify individuals and privacy was protected. Therefore, the integrated data set contained completely anonymous records.

The data are protected by laws, controls, and standards which ensure that data are held securely, that only approved people have access, and that information is used specifically for the work. Data was securely stored with NSW in the Data Linkage Centre and was inaccessible to other teams and people.

B.3.5.6 Public interest directions

In order to build and operate the NSW HSDS, a request was made to the NSW Privacy Commissioner by the Minister for Family and Community Services, for the making of two public interest directions (PIDs).

Under section 41(1) of the Privacy and Personal Information Protection (PPIP) Act and section 62(1) of the Health Records and Information Privacy (HRIP) Act, the NSW Privacy Commissioner can make a public interest direction that exempts a public sector agency from the requirement to comply with one, or more information protection principles (IPPs), or health privacy principles (HPPs), or which modifies the application of an IPP or HPP.

Section 41(3) of the PPIP Act and section 62(3) of the HRIP Act provide that the Privacy Commissioner can only make a public interest direction where satisfied that the public interest in requiring the agency to comply with the IPPs or HPPs, is outweighed by the public interest in making the direction.

B.3.5.7 Their futures matter — Public interest directions

The NSW Privacy Commissioner issued two PIDs which allowed for the creation of the HSDS.

They were signed by the NSW Privacy Commissioner after approval was given by the NSW Attorney-General and NSW Minister for Health. The Information and Privacy Commission NSW has published the Their Futures Matter PID (PIP Act) and Their Futures Matter PID (HRIP Act) on their website.

The Directions govern the extent to which TFM and participating agencies can depart from the IPPs and HPPs for the purposes of the project. They contain a clear and defined approved purpose for which data can be disclosed, collected, and used. TFM and participating agencies cannot depart from this purpose and maintain the modifications and exemptions granted. TFM reports annually to the NSW Privacy Commissioner on compliance with the Privacy Directions.

Only one copy of the HSDS is permitted to exist at any time.

B.3.5.8 Future plans for the data

Their Futures Matter will conduct an annual process to integrate new data sets and update existing data. This will create an increasingly powerful tool with clearer and more in-depth insights.

A baseline has been created, with data added each year, providing the unique ability to measure service effectiveness and population outcomes. Added data allow for more specific and detailed insights to be gained, analysing deeper into specific sub-population features and characteristics.

B.3.6 Data lifecycle stages

Data traverses through the stages of collect, transmit, store, and analyse in a Very High Control environment. Figure B.3 b) provides a simplified view of the data lifecycle. Figure B.3 c) illustrates the layers of control for data access and use, as well as data product on-sharing.

The data products created from analysis are released into a very high control or a high control environment with access limited to authorized individuals for each respective environment. This process through the layers of control, as introduced in Figure B.3 c), is illustrated in Figure B.3 d).

— Very high control environment: Raw data with personal information and sensitive information. Data and data products only accessed by those credentialled to operate in this environment. Subject to PID restrictions.

— High control environment: No access to raw data. Access only to data products which contain some personal information and sensitive location information. Only accessed by individuals independently credentialled.

Currently there is no moderate control, low control, or no control access.

Figure B.3 c) illustrates how different levels of control in the environment will have different requirements in relation to the authority to collect, use or share data, metadata about data quality and provenance, and expertise relevant to the data, project and data lifecycle stage.

A very high control environment has:

— high quality data and metadata;

— specific purpose and authority to access and use data;

— expert users experienced with the quality of the data provided and with associated metadata;

— expert analytical capability and domain expertise;

— strong governance and security at each stage of the lifecycle;

— specific restrictions on release of data and insights, or secondary use of data and insights;

— people have met general expertise requirements as well as project specific requirements and have agreed to be bound by limitations on data access and use.

Data in this environment are very high sensitivity with a very high PIF. This level of control is suitable for:

— data which can only be accessed under an external instrument such as a Public Interest Direction (PID);

— data which are reasonably personally identifiable;

— data which contain sensitive subject matter;

— data which have a well quantified quality (need not be high quality).

A high control environment has:

— explicit purpose and authority to access and use data (although might not have project specific requirements);

— expert users experienced with the data of the quality provided, and with associated metadata about data quality and provenance;

— very skilled analytical capability and domain expertise;

— strong governance and security at each stage of the lifecycle;

— explicit restrictions on release of data and insights, or secondary use of data and insights;

— people with access that have met general expertise requirements and have agreed to be bound by limitations on data access and use.

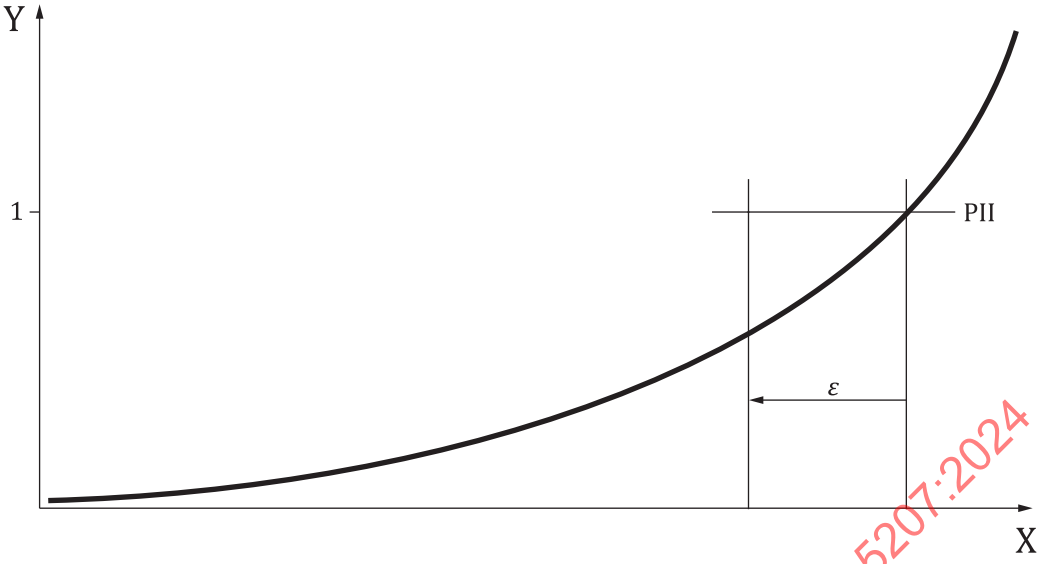
In such an environment there can be broad authority to collect, store, and use data. Data in this environment is moderately sensitive with a moderate PIF. This level of control is suitable for:

— data which are not reasonably personally identifiable;

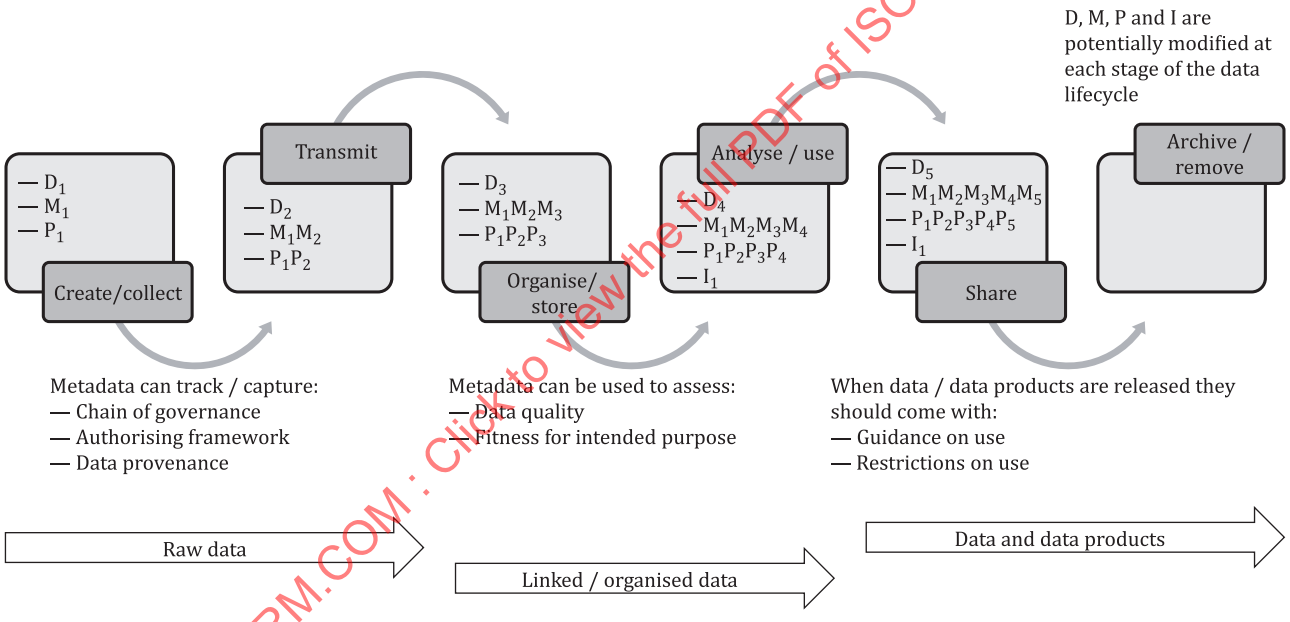
— data which contain sensitive subject matter;

— data which have a well quantified quality.

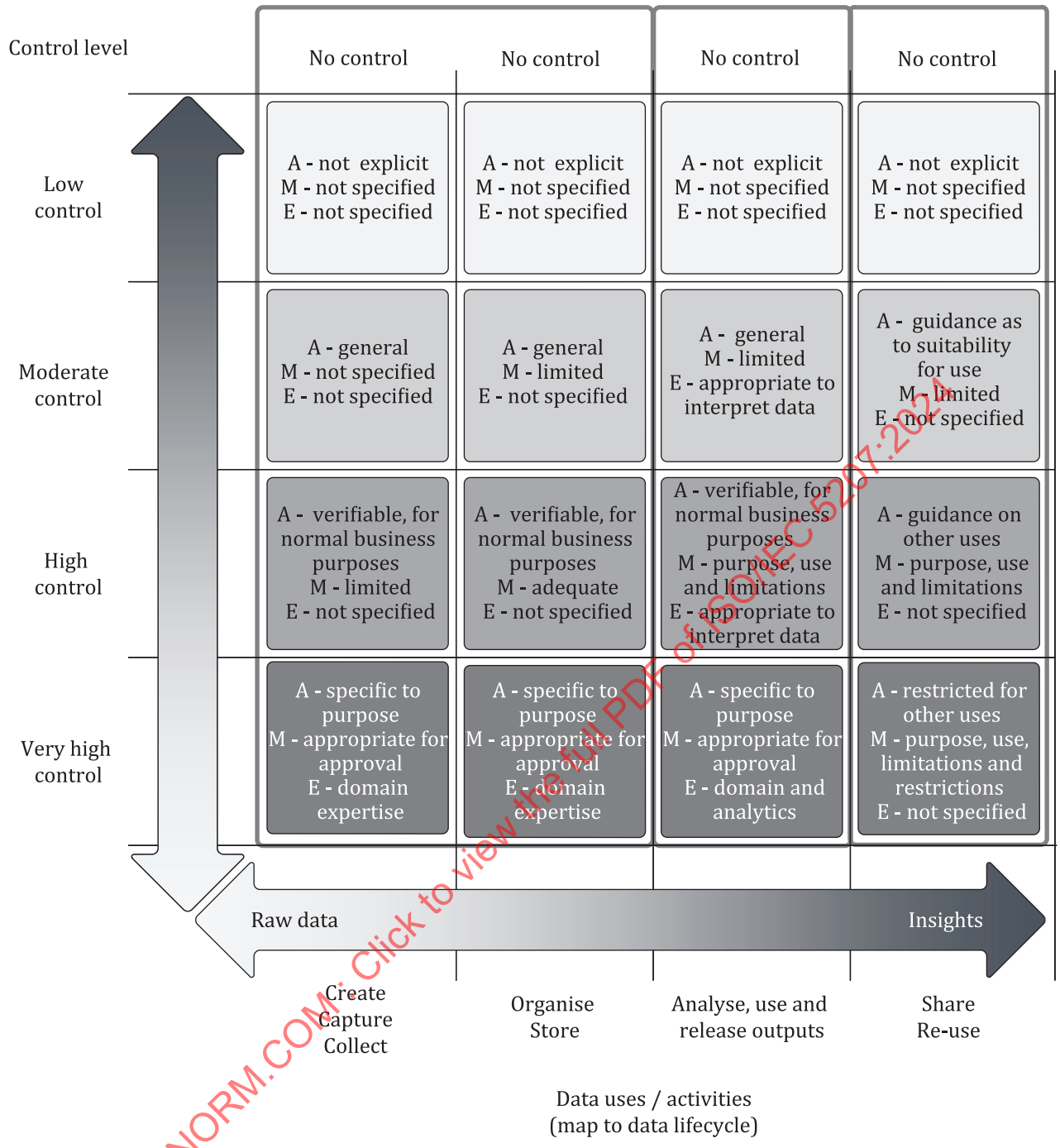
B.3.7 Figures



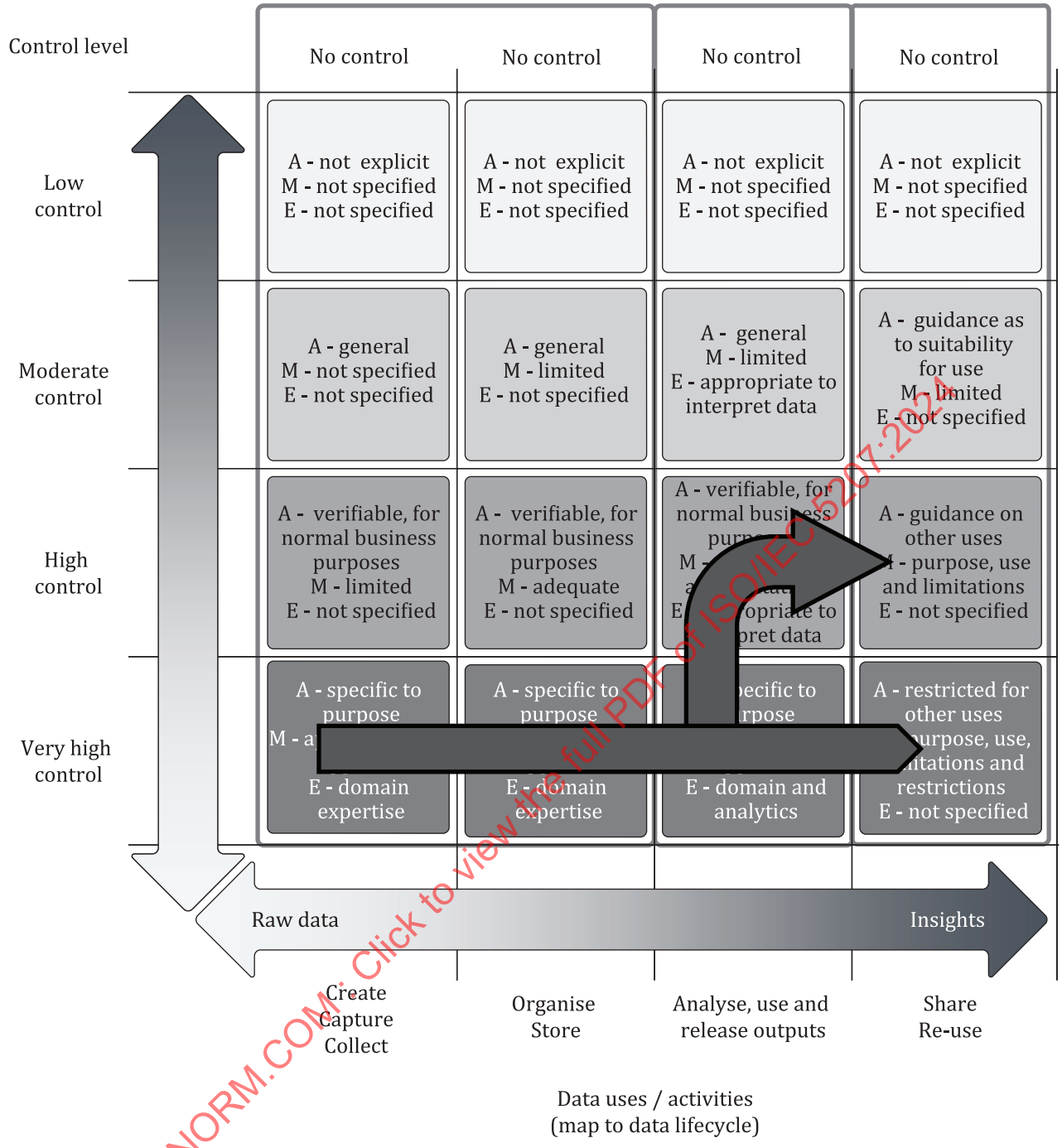
a) Concept diagram: Linked increased numbers of data sets leads to greater risk of creating Personally Identifiable Information (Data)



b) Simplified data lifecycle



c) Layers of Control for data access and use, as well as data product on-sharing



d) OOHC Data and data product access with arrows overlaying table depicted in Figure B.3 c) to show path through layers of control

Key

X	number of data sets	E	expertise
Y	personal information factor (PIF)	D	data
A	authority	P	provenance
M	metadata	I	data product

Figure B.3 — Use case 3

B.3.8 Stakeholders and stakeholder considerations

The stakeholders identified in this project included:

— Data custodians at Ministry of Health, Education, Justice, Transport and contributing agencies.

— Domain expert from contributing agencies.

— Data analysis team at NSW Data Analytics Centre (DAC).

— Cyber security and governance teams at NSW DAC.

— Policy owner at TFM responsible for outcomes.

The use case enabled agencies to make better informed decisions within their areas of responsibility. Therefore, a major aspect of stakeholder considerations related to data integrity and reliability as it was vital for decision making within each of the relevant agencies outlined in B.3.5.

B.3.9 Data characteristics

The data characteristics included:

— unit record level, de-identified data with fields of varying data quality;

— data products at aggregate (insights level).

B.3.10 Key performance indicators

KPI's came from the NSW Human Services Outcomes Framework, a set of principles-based, real-world outcomes, described in terms of key performance indicators. Analysis of the dataset allowed insights to be generated to understand barriers to achievement of indicators, or measures to positively impact performance indicators.

B.3.11 Challenges and issues

The data set exists and is being used for analysis. The challenge is increasing the scope of access (people and research questions) as well as appropriately sharing data products created with those who need to access them to act on the insights generated. While the data set remains within a highly controlled environment with limited access, the level of risk remains well managed. However, as the objective of the project is to improve service outcomes for out-of-home-care children, the project will need to consider how to balance the increase in accessibility against the privacy, safety and security of the data. This will involve a continuous assessment process of the data, control environment and the appropriateness of each level of control.

B.3.12 Societal concerns

This data asset supports United Nations Sustainable Development Goal 3 to ensure healthy lives and promote well-being for all at all ages through reform of existing systems considered to be under performing.

The major concerns are balancing release of personal and sensitive data against the Public Interest. Assessment performed by NSW Privacy Commissioner.

B.3.13 Data security, privacy and trustworthiness

The data used to create the HSDS are protected by laws and other measures that guard privacy and keep the data secure, including a comprehensive framework of security controls and standards.

B.3.14 Key insights

Given the nature and richness of the dataset created for OOHC reform, sensitivities on creation, use and analysis of the dataset are extremely high.