INTERNATIONAL STANDARD

ISO/IEC 19944-2

First edition
2022-04

Cloud computing and distributed platforms — Data flow, data categories and data use —

Part 2: Guidance on application and extensibility

Reference number
ISO/IEC 19944-2:2022(E)

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) is a worldwide federation of national standards 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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see https://patents.iec.ch).

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 Technical Committee, ISO/IEC JTC 1, *Information technology*, Subcommittee SC 38, *Cloud computing and distributed platforms.*

A list of all parts in the ISO/IEC 19944 series can be found on the ISO and IEC websites.

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

ISO/IEC 19944-1 provides a data taxonomy, data processing and use categories and other descriptive facets that can be applied to data. All aspects of ISO/IEC 19944-1 are extensible to meet the needs of diverse users. The standardized ability to categorize data, describe uses of data and apply other facets is useful in several scenarios including the application of policy to data and in describing the use of data to stakeholders.

The aim of this document is to assist users of ISO/IEC 19944-1 by providing examples and guidance for its use across several domains. Additionally, this document provides users who need to extend ISO/IEC 19944-1 with examples and guidance.

Cloud computing and distributed platforms — Data flow, data categories and data use —

Part 2: Guidance on application and extensibility

1 Scope

This document provides guidance on the application of the taxonomy and use statements from ISO/IEC 19944-1 in real world scenarios, and how to develop extensions to the data taxonomy, data processing and use categories and data use statements.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 19944-1:2020, *Cloud computing and distributed platforms — Data flow, data categories and data use — Part 1: Fundamentals*

ISO/IEC 22123-1, *Information technology — Cloud computing — Part 1: Vocabulary*

ISO/IEC 22989,[1)] *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 19944-1, ISO/IEC 22123-1, ISO/IEC 22989 and the following apply.

ISO and IEC maintain terminology 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
Internet of Things
IoT
infrastructure of interconnected entities, people, systems and information resources together with services which processes and reacts to information from the physical world and virtual world

[SOURCE: ISO/IEC 20924:2021, 3.2.1]

3.2
PII principal
natural person to whom the personally identifiable information (PII) relates

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

1) Under preparation. Stage at the time of publication: ISO/IEC FDIS 22989.

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

4 Abbreviated terms

AI	Artificial Intelligence
CSA	Cloud Service Agreement
CSC	Cloud Service Customer
CSP	Cloud Service Provider
DNN	Deep Neural Network
EV	Electric Vehicle
IaaS	Infrastructure as a Service
IoT	Internet of Things
IPR	Intellectual Property Rights
OPD	Organizational Protected Data
PaaS	Platform as a Service
PII	Personally Identifiable Information
SKU	Stock Keeping Unit

5 Presentation of ISO/IEC 19944-1

To improve transparency and guidance about data flows and data use, ISO/IEC 19944-1 names and describes the flows of data between a device and a supporting cloud service, and how to describe the use of different categories of data by the CSP.

ISO/IEC 19944-1 provides a comprehensive set of elements which can be used to:

— assign a data category to a given dataset, e.g. personally identifiable information, organizational protected data, customer content data;

— provide a set of actions applied to data, e.g. use to provide a service, to optimize it, to provide marketing information;

— define the qualifiers for the level of de-identification applied to a dataset, e.g. "identified", "anonymized", "aggregated";

— describe a use of a certain category of data for a specified purpose with a scope of its usage.

To maintain trust between the stakeholders in the ecosystem of cloud services and to meet the requirements of laws and regulations, it is necessary for service providers to be transparent about their use of the various data types that flow within the ecosystem. ISO/IEC 19944-1 also deals with organizational data and the need to treat some organizational data in particular ways in order to ensure properties such as confidentiality and integrity are maintained.

ISO/IEC 19944-1 introduces the concept of data facets, and data facets are used to extend the expressiveness of data use statements, including adding the concept of which individuals or organizations have control over data.

6 How to apply ISO/IEC 19944-1

6.1 General

This clause provides examples of ISO/IEC 19944-1 applied to several different scenarios. Readers of this document can use the examples in this clause to then apply the data categories, data use statements and other facets to their situation.

6.2 Generic eCommerce example

6.2.1 General

Modern eCommerce applications use and generate many of the data categories specified in ISO/IEC 19944-1. Additionally, the data in eCommerce applications involves other components of ISO/IEC 19944-1 including data identification qualifiers, orthogonal facets, data processing and use categories, scopes, and data use statements. eCommerce applications ordinarily involve organizational protected data such as sales volumes, pricing, customer lists and inventory data.

This clause provides descriptions and examples of how the components of ISO/IEC 19944-1 can be used in a generic retail eCommerce application.

NOTE 1 Per the definition of CSP in ISO/IEC 22123-1:2021, the operator of an eCommerce application is a CSP to its CSCs, even if the application is run on third-party services of infrastructure or platform capabilities types.

NOTE 2 Unless otherwise indicated, all references to CSC and CSP in this document refer to the party, not to the corresponding roles.

6.2.2 Customer content data

credentials: CSCs can provide data objects such as username, password, certificates and biometrics, to access eCommerce services.

financial details: CSCs can provide financial details such as credit card numbers, bank information or other payment information, to pay for purchased products and services.

6.2.3 Derived data

6.2.3.1 End user identifiable information

telemetry data: The CSP can collect data about the use of purchased products and services.

connectivity data: The CSP can collect data about the CSC's connectivity to provide the eCommerce service. For example, the CSP can use IP address information to determine the CSC's location for shipping or pickup information.

observed usage of the service capability: The CSP can collect data about the CSCs preferences and settings for the eCommerce application along with commands issued to the application (e.g. clicking the Submit button).

demographic information: The CSP can collect demographic data (e.g. age, gender) about the CSC.

profiling data: The CSP can use the various categories and instances of data to build a user profile that helps the CSP understand the CSCs interests and preferences.

content consumption data: In cases where the eCommerce application offers content, the CSP can collect data on the CSC's content consumption.

client-side browsing history: The CSP can collect client-side browsing history to help build a profile of the CSC's interests and preferences.

search commands and queries: The CSP can collect search commands and queries to improve the eCommerce application and to understand the CSC's interests and preferences.

user location: The CSP can collect the CSC's location in order to tailor offers made to the CSC.

social data: The CSP can collect the CSC's social data to better understand the CSC's interests and preferences.

6.2.3.2 Cloud service provider (CSP) data

access and authentication data: The CSP can use access and authentication data to allow CSCs to access particular aspects of the eCommerce application.

operations data: The CSP can collect data regarding operation of the eCommerce application including information about individual CSCs' use of the eCommerce application.

6.2.3.3 Account data

account or administration contact information: The CSP will ordinarily collect account data for CSCs of the eCommerce application in order to perform sales transactions.

payment instrument data: The CSP will ordinarily collect payment information from CSCs to perform sales transactions.

6.2.3.4 Organizational protected data:

price lists and pricing algorithms: The CSP can designate pricing information as organizational protected data.

sales data: The CSP can designate sales data as organizational protected data.

customer lists: The CSP can designate its customer lists as organizational protected data.

inventories: The CSP can designate its inventory data as organizational protected data.

6.2.4 Data identification qualifiers

identified data: CSPs operating eCommerce applications will ordinarily collect identified data such as account data, transaction data and profiles linked to account data that contains PII.

pseudonymized data: The CSP can substitute aliases for PII to protect individual privacy when processing of the data does not require PII but there is a potential need to link data back to specific CSCs.

unlinked pseudonymized data: The CSP can delete PII or use aliases when processing data where there is no desire to link the data back to specific CSCs.

anonymized data: The CSP can un-link and alter attributes of the data in a way that does not alter the meaning of the data but makes it reasonably impossible to identify individual CSCs directly or indirectly even if the data is combined with other data.

aggregated data: In some cases, CSPs operating eCommerce applications can aggregate transactional data for reporting and analysis where the PII in each transaction is not relevant to the task.

6.2.5 Orthogonal facets

classification: The data used or generated by an eCommerce application can be classified as High Business Impact, Medium Business Impact or Low Business Impact or using an alternate scheme that describes the significance of the data.

categorization: Data can be categorized according to what the data describes. Examples of eCommerce data categories can include CSC account information, SKU information, price information, sales transactions and inventory information.

operational control: This facet includes the basic actions that can be taken on the data. These actions ordinarily include create, read, update, delete, copy and move.

legal entity: Legal entities that can control eCommerce data include individuals, organizations and public institutions.

legal means: eCommerce data can be protected by several legal means according to local laws and regulations.

6.2.6 Data processing categories

horizontal partitioning or sharding: eCommerce data can be horizontally partitioned based on defined attributes. For example, a dataset of sales transactions can be partitioned by SKU or CSC.

vertical partitioning: eCommerce data can also be vertically partitioned by keeping only a subset of attributes. For example, an inventory dataset can be vertically partitioned by removing the item weight attribute.

data association: eCommerce data can be stored in different datasets which are then linked together. For example, a dataset of transactions can link to other datasets that contain CSC information and to datasets that have detailed information about the products or services purchased.

data aggregation/consolidation: Data analysis can require that different eCommerce datasets be aggregated or consolidated. For example, sales transactions from different geographies can be stored in different datasets which are then aggregated to see global sales information.

data accumulation: Datasets containing eCommerce data ordinarily accumulate new records over time. For example, a dataset containing eCommerce transactions can grow by millions of records each day. These datasets can then be used for time-series analysis to identify trends.

data fusion: eCommerce data from multiple datasets can be combined and then reduced to obtain an improved dataset.

data standardization: Entries in eCommerce datasets can include data recorded in the wrong field. For example, the first and last name of a CSC recorded in a field labelled “First Name” or a postal code entered in a filed labelled “Street Address”. Data standardization means placing the data entries into the correct fields across the entire dataset.

data validation and correction: Data in an eCommerce dataset can be the wrong type or format or simply be incorrect. For example, if an entry of “Yes” is in a field labelled “Age”, the entry is probably incorrect. If an entry of “2020” is in an “Age” field, it can be the correct type and format but is clearly wrong. The process of data validation and correction is the process of correcting the type, format and values of data entries.

data enrichment: Entries in eCommerce datasets can be missing which can interfere with transaction calculations or analysis algorithms. For example, if the price field is blank for a purchase, the amount charged for the sale will be incorrect as will any downstream analysis of the dataset. Data enrichment means filling in missing data entries using some imputation process.

encryption: eCommerce data can be encrypted in flight and at rest to prevent unauthorized access and use.

replication: Copies of eCommerce data can be made and stored in different locations for the purpose of business continuity or recovery in the event of a disaster.

data deletion: CSPs operating eCommerce applications can delete data from time to time. Data deletion is often done when the data is no longer relevant to the CSP, CSC or other interested parties.

secure data deletion: In some cases, the CSP can delete eCommerce data in a way that prevents any party from ever recovering it.

re-identification: An eCommerce CSP can have the need to re-identify data that had previously been de-identified.

6.2.7 Data use categories

provide: eCommerce CSPs will ordinarily use many data categories from multiple scopes to provide the eCommerce service to its CSCs. For example, account data can be used to populate the billing and shipping address fields of a purchase transaction. The postal code portion of account data can be used to calculate shipping costs and time.

improve: Data can be used to improve the eCommerce service. For example, the postal code portion of account data can be used to identify clusters of CSCs which can then be used to site warehouses and reduce shipping times. Data from failed transactions can be used to improve the purchase process and in turn improve CSC satisfaction.

personalize: Data categories such as observed usage of the service capability, demographic information, profiling data, content consumption data and search commands and queries can be used to personalize the eCommerce service for a CSC or group of CSCs. For example, if a CSC or group of CSCs regularly use a search function to find products in a particular category, the service's home page can feature products that correlate to the CSC's search history.

offer upgrades or upsell: eCommerce CSPs can use transaction histories and profiling data to offer upgrades or upsells during the purchase transaction process or through other means. For example, if a CSC selects a quarter inch drill bit, the CSP can offer to upsell a drill bit set.

market/advertise/promote: Similar to "offer upgrades or upsell", transaction data and profile data can be used to promote specified products to a CSC or groups of CSCs.

promote based on contextual information: eCommerce CSPs can use data based on the use of the current capability or on the services and application scope to promote products to the CSC. This data use category does not make use of the CSC's previous use of the service.

promote based on personalization: The eCommerce CSP can use data to change the content of a promotion for a CSC or group of CSCs.

share: eCommerce CSPs can share data with third parties to operate the service. For example, the CSP can share the postal code portion of account data with a third-party shipper to get a shipping cost and delivery time estimate. The CSP can share data with third parties when outsourcing functions such as analytics.

share when required to provide the service: The CSP can be required by law, or by contractual obligations to share eCommerce data.

collect: eCommerce CSPs can collect data from many different categories such as financial details, observed usage of the service, demographic information, profiling data, content consumption data and others. Additionally, the CSP will ordinarily store, prepare and pre-process the collected data for down-stream uses including AI.

train: eCommerce CSPs can use training data from one or more categories to train machine learning models in AI systems. For example, the CSP can use data from the observed usage of the service capability to train a product recommendation model that is presented to CSCs during their use of the eCommerce system.

6.2.8 Scopes

6.2.8.1 General

As described in ISO/IEC 19944-1, scopes help the CSC to understand the use of data in terms of where the data comes from (source scope), where it is used (use scope) and what the result of the use is (result scope). A scope is ordinarily used as a part of a data use statement.

eCommerce applications can be composed from multiple, uniquely identifiable services such as product inventory, product search, product recommendations, product reviews, order entry, shipping and payment. As such, scopes can span a single service, groups of services or the entire eCommerce application.

6.2.8.2 Source scope

Data can be sourced from several components of the eCommerce application, particularly those where CSCs enter information or make selections. Examples of source scopes can include order entry, product search and product reviews.

The service as described in the CSA can also be used to describe the source scope.

6.2.8.3 Use scope

The eCommerce application itself can be the use scope. In some cases, the data can be used in an ancillary or third-party service. For example, data from product reviews can be sent to the product manufacturer as input to future product improvements. Order input information along with account data can be sent to a third-party service to calculate sales tax or shipping costs.

6.2.8.4 Result scope

Result scopes are an application of the data use categories. In the case of an eCommerce application, examples of result scopes include providing data, improving or modification of the data, and personalization of the data

6.2.9 Data use statements

6.2.9.1 General

As described in ISO/IEC 19944-1, data use statements are composed of data categories, actions, source scopes, use scopes and result scopes to provide a complete description of where data comes from, how it is used and for what purpose. Additionally, a data use statement can include a data identification qualifier.

6.2.9.2 Examples of data use statements

The following are examples of data use statements that can be used in conjunction with an eCommerce application.

EXAMPLE 1

“Anonymized telemetry data collected by the product monitoring service is used by the product manufacturer to improve the product”.

In EXAMPLE 1, the data category is “telemetry data”, “collected’ is the action, the source scope is “the product monitoring service”, the use scope is “the product manufacturer” and the result scope is “improve the product”. “Anonymized” is a data identification qualifier that tells the CSC that their personally identifiable information has been unlinked.

EXAMPLE 2

"Aggregated, pseudonymized content consumption data is used to train a model that is used to improve product recommendations."

In EXAMPLE 2, there are two data identification qualifiers 1) aggregation and 2) pseudonymization. "Content consumption data" is the data category, "training a model" is the action and "improve product recommendations" is the result scope. Note that, in this case, there is no explicit source or use scope.

EXAMPLE 3

"All data collected and used by this eCommerce service is encrypted during transport and storage to improve the security of this service".

In EXAMPLE 3, the data category is "all data", the source scope is "this eCommerce service", the action is "encrypt", the use scope is "transport and storage" and the result scope is "improve the security of the service".

EXAMPLE 4

"Identified, profiling data collected by this eCommerce service is used by the product recommendation service to provide, personalize and improve product recommendations."

In EXAMPLE 4, "profiling data" is the category, "this eCommerce service" is the source scope, actions include "provide and improve" and "product recommendations" is the result scope. The data identification qualifier of "identified" lets the CSC know that their PII can be linked to the data used in this instance.

EXAMPLE 5

"Location data collected by the service described in the service agreement will be used by the marketing service to offer upgraded products and services".

In EXAMPLE 5, the source scope is "the service described in the service agreement" rather than being explicitly identified. "Location data" is the data category, "marketing service" is the use scope, "offer" is the action and "upgraded products and services" is the result.

6.3 Privacy examples

6.3.1 General

ISO/IEC 27701[4] refers to ISO/IEC 19944-1 for two uses:

— Providing terms for describing the purpose of the processing of PII.

— Providing definitions of data identification qualifiers that can be used to classify the degree to which the data can identify a PII principal or associate a PII principal with a set of characteristics in the PII.

This clause provides examples of how to utilize ISO/IEC 19944-1 to fulfil these uses.

6.3.2 Describing the purpose of the processing of PII

The data categories, data processing categories and data use categories in ISO/IEC 19944-1 can be utilized to describe the purpose of the processing of PII in a consistent way.

EXAMPLE 1 shows a simple statement that explains how PII is used to provide a service.

EXAMPLE 1

"Your credentials will be used by the service to provide the service."

In EXAMPLE 1, the data category is "credentials" (a sub-category of customer content data), "the service" is both the use scope and the result scope, and "provide" is the action.

Credentials are ordinarily required to access a service or resource and can include usernames, passwords and biometric data.

EXAMPLE 2 shows a compound statement of how a CSP can protect and use PII.

EXAMPLE 2

"Your credentials and financial details will be encrypted at rest and in motion and used by the service to provide the service."

In EXAMPLE 2, "encryption" is an additional action and "financial details" is an additional category. Credentials and financial details are often protected through encryption to prevent unauthorized use in the event of a security breach.

EXAMPLE 3 shows how a CSP can specify multiple uses for several related data categories along with a source scope.

EXAMPLE 3

"Demographic information, profiling data, content consumption data, search commands and queries, CSC location and social data collected by the service will be used by the service to provide, improve and personalize the service."

In EXAMPLE 3, "data collected by the service" is the source scope.

6.3.3 Using data identification qualifiers with PII

The data identification qualifiers defined in ISO/IEC 19944-1 can be used to explain which PII minimization techniques have been applied. In general, data identification qualifiers can be added to most data use statements.

EXAMPLE 1 shows how the data identification qualifier "identified data" can be used with 6.3.2, EXAMPLE 1.

EXAMPLE 1

"Your credentials, consisting of identified data, will be used by the service to provide the service."

EXAMPLE 2 shows how the data identification qualifier "anonymized data" can be used with elements of 6.3.2, EXAMPLE 3.

EXAMPLE 2

"Anonymized demographic information, profiling data, content consumption data, search commands and queries, CSC location and social data collected by the service will be used by the service to improve the service."

6.4 Organization identifiable data examples

6.4.1 General

ISO/IEC 22624[2] describes a framework for the structured expression of data-related policies and focuses on PII. These examples will address the application of ISO/IEC 19944-1 to organization identifiable data that is not PII (i.e. non-personal data).

Governments have instituted regulations and policies on organization identifiable data often arranged by industry sector. Regulations and policies for organization identifiable data often apply to PII. Governments are now considering regulation and policy for non-personal organization identifiable data to address national and societal interests. Examples of emerging regulations and policies include requirements to store and process organization identifiable data within a specified jurisdiction and requirements to share specified categories of data with specified stakeholders. Examples of data categories that governments consider as non-personal include inferences, geospatial, earth observation,

environmental, meteorological, transport and measurements obtained from machines and physical systems.

De-identified data can be considered non-personal data as well, provided it cannot be linked back to a person. For information about using ISO/IEC 19944-1 with de-identified data see ISO/IEC 22624[2].

NOTE The examples used in this subclause include categories not defined in ISO/IEC 19944-1 but are examples of how ISO/IEC 19944-1 can be extended as described in Clause 7 of this document.

6.4.2 Organization identifiable data location requirement examples

Table 1 shows how data categories, data processing categories, scopes and other data facets from ISO/IEC 19944-1 can be used to express data location policies or requirements. These elements can be used by any type of organization (e.g. government, private sector, universities).

Table 1 — Example expression of geolocation policy/requirement for organization identifiable data

Data category		Geolocation policy/requirement		Jurisdiction
Primary	**Secondary**	**Storage**	**Processing**	
Organization identifiable	National security	Locally	Locally	National
Organization identifiable	Electric grid data	Locally	Locally	National
Organization identifiable	Meteorological data	Global	Global	National/State/Local
Organization identifiable	Employment statistics	Locally	Locally	National

The contents of Table 1 can also be expressed as follows:

EXAMPLE 1

"You must store and process national security data originating in [Country] within the territorial boundaries of [Country]."

EXAMPLE 2

"You must store and process data related to the operation of electrical grids originating in [Country] within the territorial boundaries of [Country]."

EXAMPLE 3

"You may store and process meteorological data in any country subject to local, state or national regulations."

EXAMPLE 4

"You must store employment statistics data originating in [Country] within the territorial boundaries of [Country]."

6.4.3 Organization identifiable data sharing requirement examples

Table 2 shows how data categories, use categories and scopes can be used to express data sharing policies and requirements.

Table 2 — Example expression of data sharing policies/requirements for organization identifiable data

Data category		Sharing required?	Source scope	Use scope	Use category	Result scope
Primary	Secondary					
Organization identifiable	Machine telemetry	Yes	Farm equipment	Farm equipment owners	Improve	Farm equipment operation, Farm yields
Organization identifiable	Environmental	Yes	Farm soil, water	Farm owners	Improve	Farm yields
Organization identifiable	Machine telemetry	Yes	Automobiles	Automobile repair facilities and owners	Provide, Improve	Automobile repair and maintenance
Organization identifiable	Meteorological	No	Meteorological measurements, observations and predictions	Any	Provide, improve, share, collect, train	Products and services, research, weather prediction
Organization identifiable	Employment statistics	Yes	Employment data	As allowed or required by law	Improve, provide	Economy, employment assistance

The contents of Table 2 can be expressed as follows:

EXAMPLE 1

"You must share the machine telemetry data collected from farm Equipment with farm equipment owners who may use the data to improve the operation of farm equipment and farm yields."

EXAMPLE 2

"You must share environmental data collected from farm soil and water with farm owners who can use the data to improve farm yields."

EXAMPLE 3

"You must share the machine telemetry data collected from automobiles with automobile repair facilities and automobile owners who can use the data to repair and maintain automobiles."

EXAMPLE 4

"You may share meteorological data collected from meteorological measurements, observations and predictions with any party who may use the data to provide and improve products and services, conduct research and make weather predictions."

EXAMPLE 5

"You must share employment statistics derived from employment data with any party authorized by law who can use the data to improve the economy and provide employment assistance."

6.5 AI example

6.5.1 General

Artificial intelligence and machine learning are data-intensive applications that can benefit from the use of a standardized data taxonomy and data use statement expressions. In the development and use of machine learning models, explainability and transparency can be necessary (e.g. where the training data used for training of the machine learning model originated, who controls or has access to it, the nature of the data, and the methods for training).

It would be helpful to first consider how data is handled in a machine learning application. It is first acquired or collected, then pre-processed and prepared for use in the model training. Once the appropriate algorithm and learning approach is selected for the given application, the prepared data is used to train the selected algorithm It can be necessary to repeat this process multiple times until the desired model is sufficiently trained for use in the AI system.

Figure 1 shows the stages of data processing described in 6.5.1.

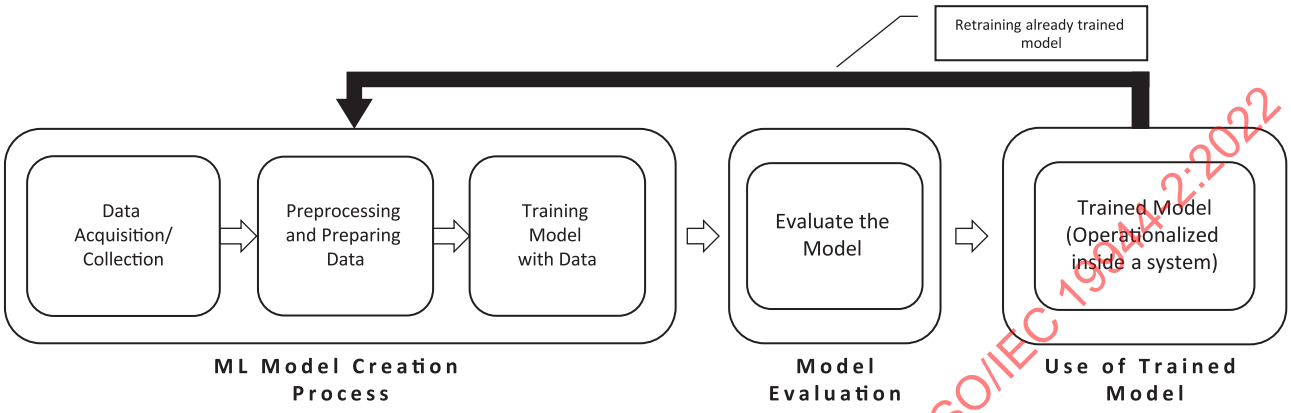

Figure 1 — Example of a data acquisition and processing pipeline for AI

To provide transparency and explainability of the artificial intelligence applications one can explain the types of data that were involved in the development of the model. The multi-faceted taxonomy introduced in ISO/IEC 19944-1 is well suited for this purpose. At every stage of the machine learning development process, the data involved can be described using some or all the facets depicted in ISO/IEC 19944-1:2020, Table 1, *Commonly known, orthogonal facets of data*. Use of the taxonomy enables data use statements that can provide for transparency and explainability of an AI process by showing what was used to train the machine learning models, and how.

6.5.2 Facial recognition — Privacy-centric AI example

Consider a machine learning application that is fed a video stream from a camera installed at an airport terminal. Assume the application is trained to detect patterns in a passengers' face and body that can fit a profile of a known drug trafficker or a terrorist set to hijack airliners. Such an AI system can be built as a customized version of a general-purpose facial recognition software that is specifically trained with photos of known terrorists or drug traffickers, perhaps using DNNs.

Given the nature of the data used, this is a privacy centric scenario where there can be a need to explain where the photos of subjects used in model training came from, and how they were used in the training. There is also a need to explain why or how the customized facial recognition algorithm does not inadvertently discriminate against minority groups, such as people with certain facial complexion, body shape or posture that can map to certain racial or national profiles.

Figure 2 describes how the "categorization" facet of data defined in ISO/IEC 19944-1 can be used to provide a precise data use statement about the development of the airport security facial recognition application:

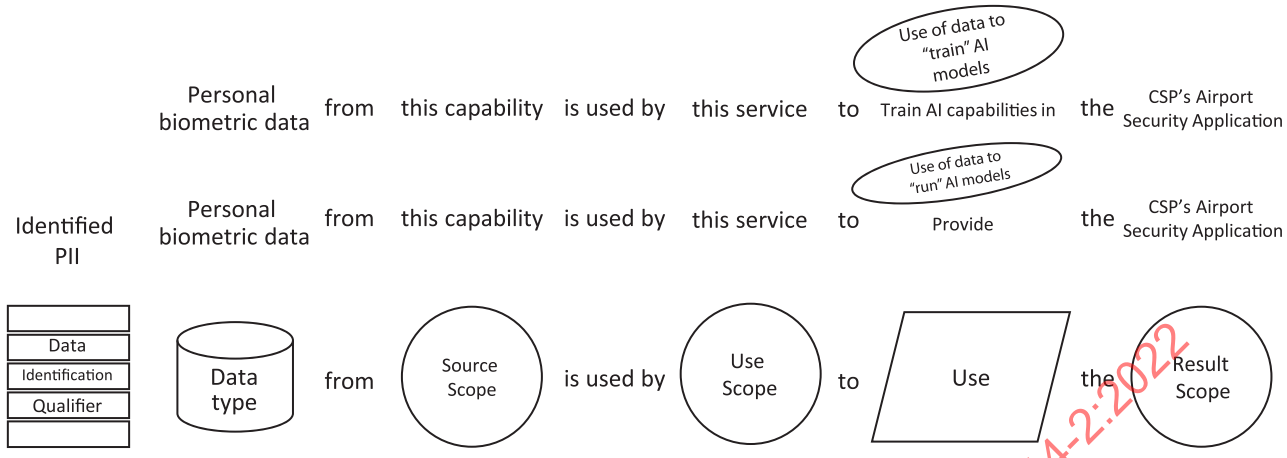

Figure 2 — Data use statement for PII — Training and use of AI facial recognition model for an airport security application

The data use statements can be more specific as to the nature of the data that was used to customize the general-purpose facial recognition used. In this example, as shown in Figure 2, the facial biometric data used in customization can come from classified law enforcement databases for known criminal or terrorist groups. In such a case, the extensibility mechanism for the data taxonomy in ISO/IEC 19944-1 can be used to make more precise, and scenario specific data use statements, as shown in Figure 3.

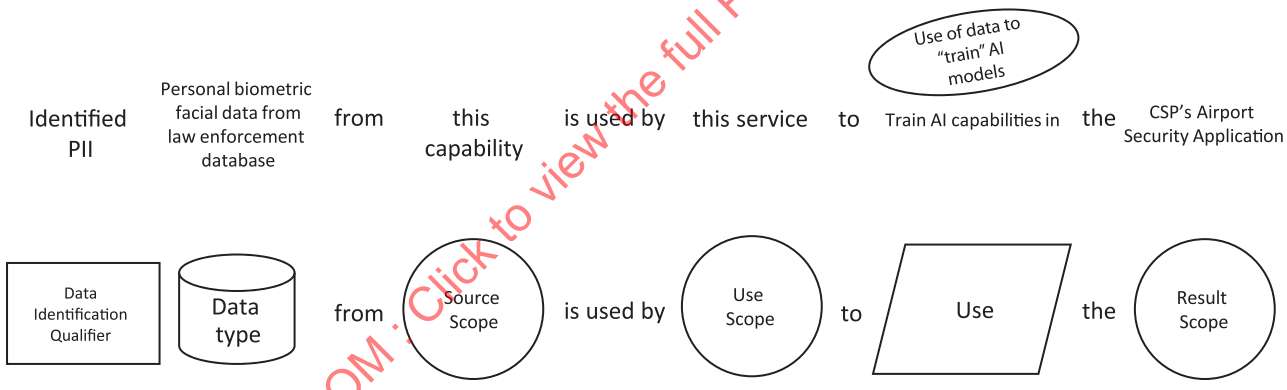

Figure 3 — Data use statement for PII — Training of a facial recognition model for an airport security application using law enforcement data without de-identification

In Figure 3, facial data obtained from a law enforcement database about known terrorists or criminals are used, without de-identification, to custom-train the general purpose DNN facial recognitions model. The data use statement in Figure 3 is precise in disclosing that the training data was obtained from law enforcement database to perform additional, customized training of the facial recognition model. Furthermore, the data use statement makes it clear the facial data was used without de-identification (presumably to maximize the impact of the data used in training of the model).

The extensibility used in Figure 3 is sub-typing an existing sub-category of data, namely the personal biometric data (ISO/IEC 19944-1:2020, 8.2.2.6). The sub-type in this example, the facial data obtained from a law enforcement database, is used to provide more granular, scenario-specific data use statement.

6.6 IoT example

6.6.1 General

This example involves commercial scenarios where company confidential data or intellectual property as captured in electric vehicle (EV) telemetry. These data are used to train machine learning models that can predict maintenance issues of those EVs in advance, lowering cost as well as improving safety.

6.6.2 Electrical vehicles

EVs are designed with numerous digital sensors that are continually collecting telemetry data from various components inside the car. These data are processed on-board in real time and can be communicated to the manufacturer's cloud services for additional processing and analytics in cloud computing data centres, where the data can then be used to train machine learning models for predictive maintenance or even traffic flow management.

The EV contain a set of IoT devices for the purpose of on-board data collection and communication with the edge computing points of presence or cloud services (see ISO/IEC TR 23188[3] for the distributed computing environment that consists of cloud computing, edge computing and IoT devices). The on-board IoT devices collect and transmit data critical to the health and operation of the EV. The same data can later be analysed by cloud computing applications and be used to train AI models for various purposes, including predictive maintenance or predictive traffic flow management and control.

Such AI algorithms are often costly to train and develop and requires highly specialized data science expertise. Therefore, it can make sense for several EV companies to pull their resources together, or jointly outsource the task to a third party. In such a case, the EV telemetry data, containing potentially confidential data or intellectual property in need of protection, can be merged to form shared pools of data.

Given that the data extracted from EV telemetry contains manufacturer intellectual property data that can be protected (e.g. the data can contain clues to the key competitive EV design and architecture), the shared EV telemetry data used in model training can need de-identification to protect against inadvertent sharing of confidential corporate intellectual property.

The following data use statements are examples for use in commercial contracts where the EV manufacturers and the predictive maintenance developers agree to share EV telemetry data. The added precision obtained by use of a common, standardized taxonomy provides for additional confidence and assurance that that shared data is used only for the specific agreed purpose, and that no unintended loss of intellectual property would happen, given the de-identification of data performed.

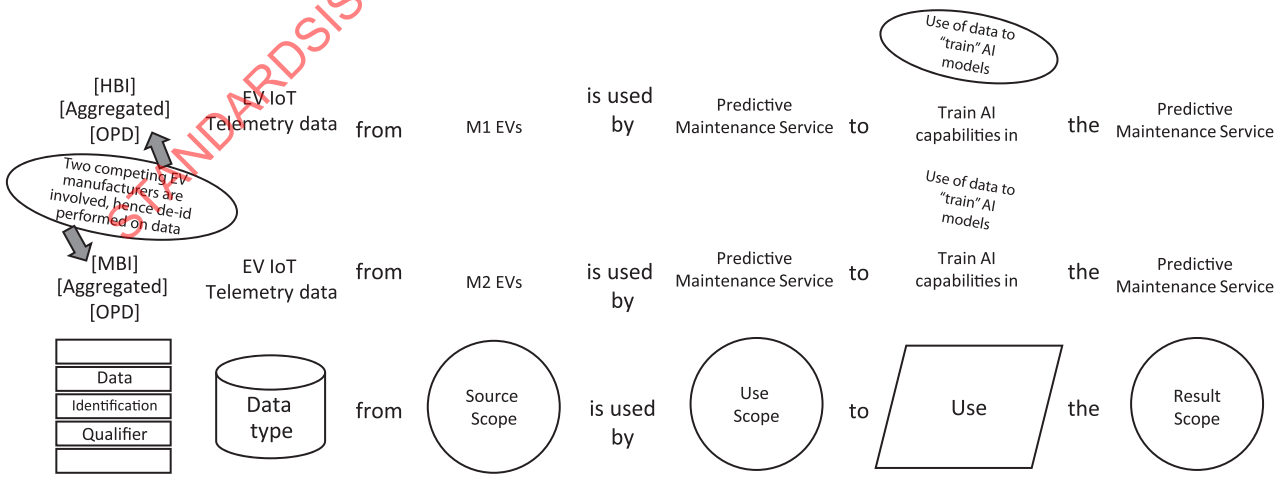

Figure 4 — Use statement for OPD — Training of an AI model for electric vehicle predictive maintenance service, using onboard IoT telemetry data

In Figure 4, two fictitious EV manufacturers, Manufacturer 1 (M1) and Manufacturer 2 (M2), have entered into a data sharing agreement, where they pool the telemetry data collected from the IoT devices onboard their sold cars on the roads to form a shared depot of EV data. They have then entered an agreement with a developer of AI solutions for predictive maintenance or traffic flow management and control. The developer has obtained access to the shared pool of IoT telemetry data and will use them to train their models for predictive maintenance or traffic flow management.

Given M1 and M2 are competitors in the EV market, they do not wish to share telemetry data that can be deciphered by their competitor to uncover trade secrets in EV design. Therefore, before they share their IoT telemetry data, they aggregate certain parts of the data that is unrelated to the maintenance aspects of the cars, or independent of use by traffic flow models.

The two examples of data use statements in Figure 4 provide precise disclosure about what data has been used in training of the AI models. Such precision disclosures can be useful in composing data sharing agreements thereby empowering the development of commercial contracts for data sharing for AI system development.

Notice that the OPD attribute is used to designate the EV IoT data; this indicates that the EV data is under the control of an organization, and the control is derived from, and based on known laws/ regulations or a contract.

The disclosed data sub-type, namely EV IoT Telemetry data, is an example for extending customer content data (see ISO/IEC 19944-1:2020, 8.2.2). In this example, EV IoT Telemetry data is used as a custom sub-type of customer content data, and not telemetry data (ISO/IEC 19944-1:2020, 8.2.3.2.2). The reason is that the latter is a sub-type of derived data, whereas the custom data type introduced here for this example needs to be a sub-type of customer data.

7 How to extend ISO/IEC 19944-1

7.1 General

To make ISO/IEC 19944-1 useful across a wide range of use cases, the data taxonomy, custom data facets, data processing categories and data use categories may be extended. This clause provides guidelines and examples to help users of ISO/IEC 19944-1 tailor extensions to their applications.

7.2 Data taxonomy

7.2.1 General

ISO/IEC 19944-1 states that the defined data taxonomy is intended to be extensible so that CSPs can define new data sub-types to suit the needs of their cloud services. This clause provides guidelines for creating new data sub-types (extensions) along with examples.

7.2.2 Guidelines for extending the data categories defined in ISO/IEC 19944-1

Whenever possible, CSPs should use the data categories defined in ISO/IEC 19944-1.

CSPs should use the highest possible level of abstraction of a data category. This prevents a proliferation of data categories and makes it easier for stakeholders to understand how a category of data is processed and used. A new sub-type should only be defined if there is a need to apply data processing or data use statements that cannot be made about the parent data category.

A new sub-type should connect to the hierarchy of data categories defined in ISO/IEC 19944-1.

At a minimum, a new sub-type should connect to one of the four top-level data categories defined in ISO/IEC 19944-1 (customer content data, derived data, CSP data and account data). A new sub-type should not be added at the same level as the four top-level categories.

A new sub-type should be clearly described to stakeholders along with its relationship to other categories. Descriptions of existing data categories are provided in ISO/IEC 19944-1.

If a new sub-type is created, any data processing or use statement applied to a higher-level category will also apply to the sub-type. However, a sub-type can be explicitly excluded from a data processing or use statement.

7.2.3 Example of extending the cloud service provider (CSP) and customer content data categories

An eCommerce CSP can create an "inventory" data category. Noting the guidelines in 7.2.2 and examining the existing data categories and sub-types in ISO/IEC 19944-1 suggests that "inventory" can fit under the CSP data category and the Customer content data category but does not appear to fit into the existing sub-types. Therefore, it would be reasonable to create a new sub-type for inventory:

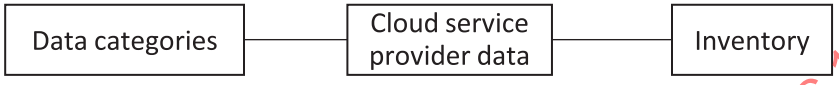

Figure 5 — Extension of the cloud service provider (CSP) data category

Figure 5 shows that inventory is a sub-type of CSP data when a data processing or data use statement is made between the eCommerce CSP and its retail CSCs or suppliers.

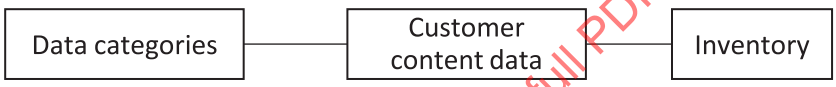

Figure 6 — Extension of the customer content data category

Figure 6 shows that inventory is a sub-type of customer content data when a data processing or data use statement is made between the IaaS or PaaS CSP and the eCommerce CSP.

A description of the inventory sub-type can be as follows:

"Inventory is a sub-type of the CSP information (or customer content) data category that covers all physical goods in the company's legal possession that are intended for resale. Inventory data can include items such as manufacturer name, model number, serial number, UPS number, SKU number, ISBN number, package dimensions, package weight and quantity."

7.2.4 Example of extending the demographic information sub-type

As shown in Figure 7, an eCommerce CSP can create sub-types of the demographic information data category for items such as CSC age, location and income level. In this case the extensions follow the hierarchy defined in ISO/IEC 19944-1.

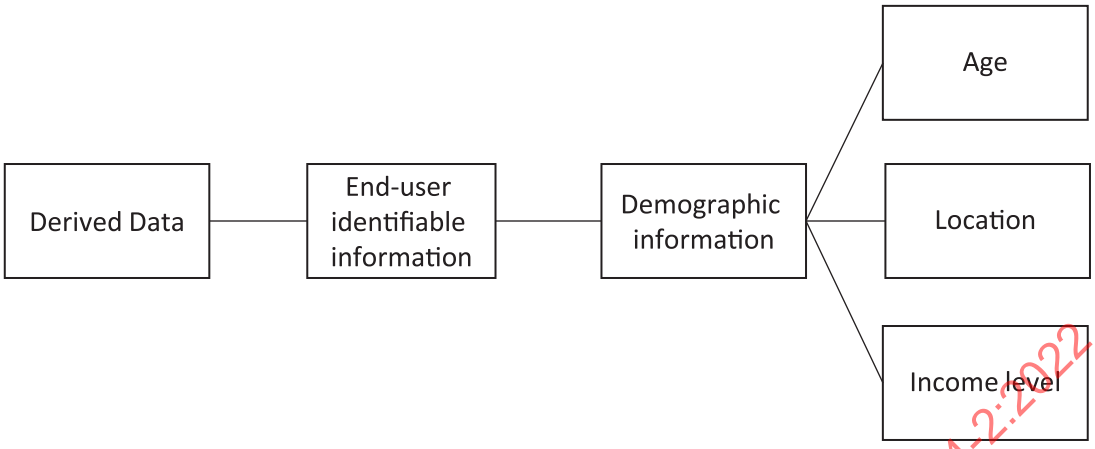

Figure 7 — Extension of the demographic information data category.

Descriptions of these new sub-types can be as follows:

— "Age is a sub-type of the demographic information data category that is intended to capture the specific age or age range of a CSC."

— "Location is a sub-type of the demographic information data category that is intended to capture the CSC's postal code, latitude and longitude or other location information."

— "Income level is a sub-type of the demographic information data category that is intended to capture the CSC's income as a categorial range."

As stated previously, any data processing or use statement made about a high-level category will automatically apply to all lower levels unless an explicit exclusion is made.

7.2.5 Example of extending the financial details sub-type

As shown in Figure 8, an eCommerce CSP can create a sub-type of the financial details data category for credit scores in the event the sub-type requires specific data processing or data use statements.

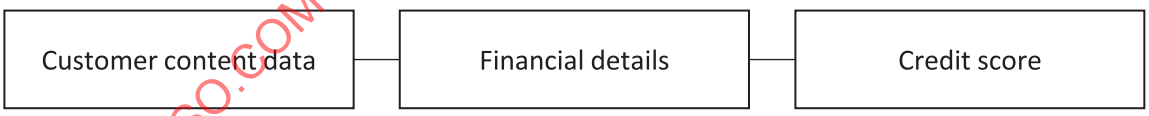

Figure 8 — Extension of the financial details data category

A description of the credit score sub-type can be:

"Credit score is a sub-type of the financial details data category and is intended to capture the CSC's credit score as provided by a third-party credit agency."

7.3 Custom data facets

7.3.1 General

ISO/IEC 19944-1 describes orthogonal data facets and allows for custom data facets. This clause provides guidance and an example for constructing custom facets.

7.3.2 Guidance on creating custom data facets

— A custom data facet should be orthogonal to any other custom or defined data facet.

— A custom hierarchy of data facets, including custom facets, can be developed if needed.

— A custom facet can be designed from the perspective of a group of stakeholders.

7.3.3 Example custom data facet

An eCommerce CSP can create facets for their data along the lines of traditional data management terminology. For example, the organization can additionally categorize data as master data or transaction data. In this case, both master data and transaction data can be composed of customer content data (e.g. name, address, order details) and CSP data (e.g. inventory, price lists, tax rates)

7.4 Data processing

7.4.1 General

This clause provides guidelines and examples for extending the data processing categories in ISO/IEC 19944-1.

7.4.2 Guidelines for extending data processing categories

Whenever possible, CSPs should use the data processing categories defined in ISO/IEC 19944-1:2020, 9.2.

7.4.3 Examples for extending data processing categories

EXAMPLE 1

This example, as shown in Figure 9, describes how the replication data processing category can be extended. Data replication can be used to address disaster recovery and business continuity requirements along with other events that can cause data loss. Creating mirror copies of data is one specific method for replication and can be defined as a data processing category.

Figure 9 — Extension of the replication data processing category

A description of the mirroring data processing category can be:

“Mirroring is a sub-type of the replication data processing category and is intended to indicate that more than one copy of the data is stored in geographically diverse locations.”

A data use statement that includes the mirroring data processing category can be:

“Customer content data is mirrored in at least two geographically diverse locations.”

EXAMPLE 2

This example describes how a new top-level data processing category can be created.

Data storage can be organized as tiers depending on the CSC’s latency requirements. Hot, cool and cold are colloquial terms used to describe storage tiers where the hot tier has the lowest latency and the cold tier has the longest latency and can be used for data archiving.

Storage tier deprecation can be a new top-level data processing category.

A description of the storage tier deprecation data processing category can be: