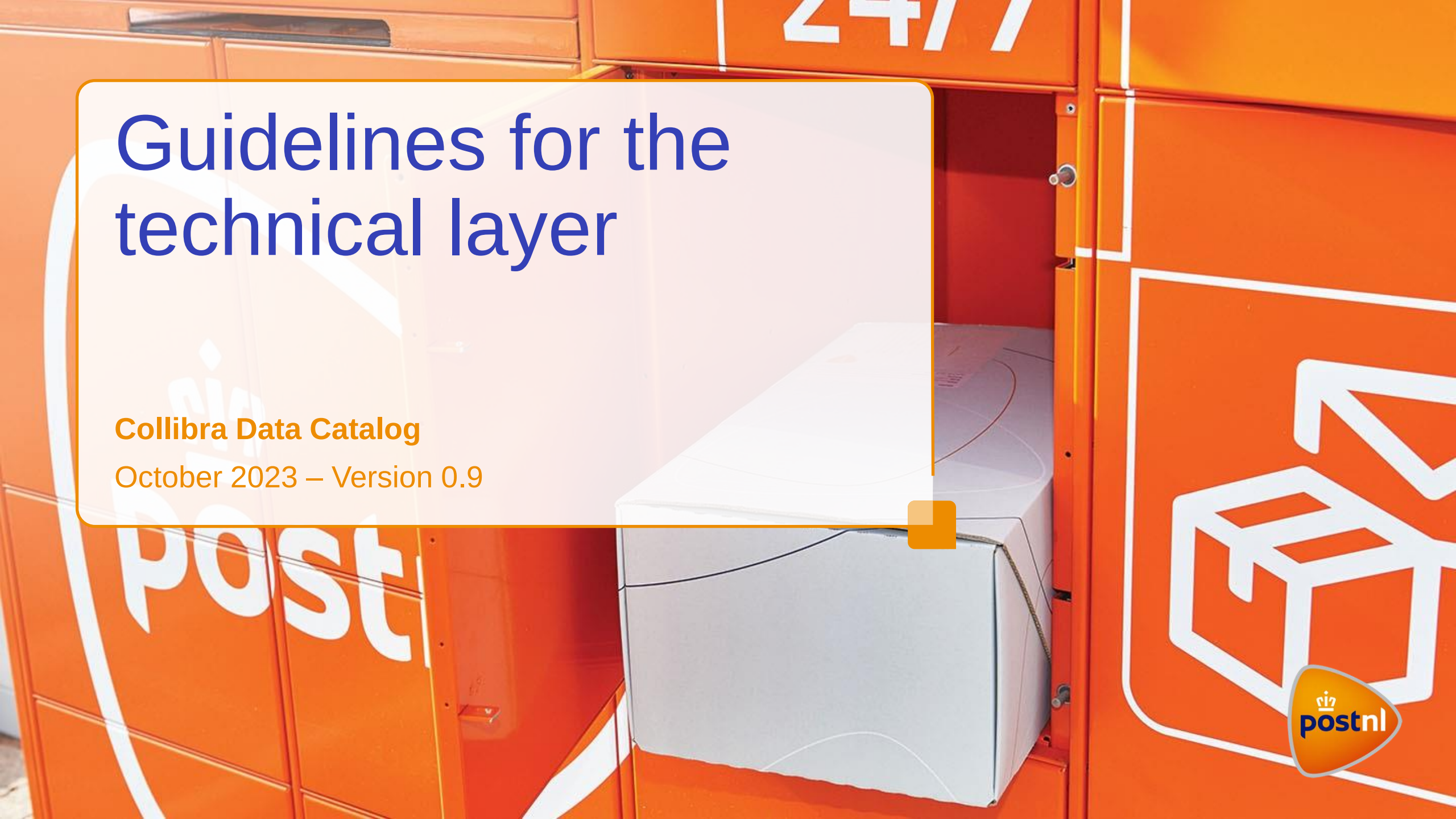
Guidelines for the technical layer

Collibra Data Catalog

October 2023 – Version 0.9



Index

Introduction

The PostNL Data Catalog

Introduction

PostNL uses the Collibra Data Governance Center (DGC) platform to support its data organization in the areas of data quality, data governance, and data analysis. It assists in locating data, provides metadata inventory, and offers information necessary to determine if the data is suitable for its intended use.

The PostNL Data Catalog is a tool that enables PostNL's data organization to enhance the accessibility, accuracy, and relevance of data across the entire company. This provides crucial support for:

- **Data Usage:** Metadata enhances knowledge about data. The more users know about data, the better they can determine its usability and limitations.
- **Data Management:** The data catalog provides insights and a better understanding of the data that PostNL possesses. This makes the data known and manageable and is a prerequisite for the professionalization of data management capabilities such as data governance and data quality management.

The PostNL Data Catalog is one of the solution building blocks within the data capability 'Meta Data Management.'

*"**Metadata Management** is the discipline that involves the collection, maintenance, and standardization of metadata (including access and distribution)."*

Structure of the PostNL Data Catalog

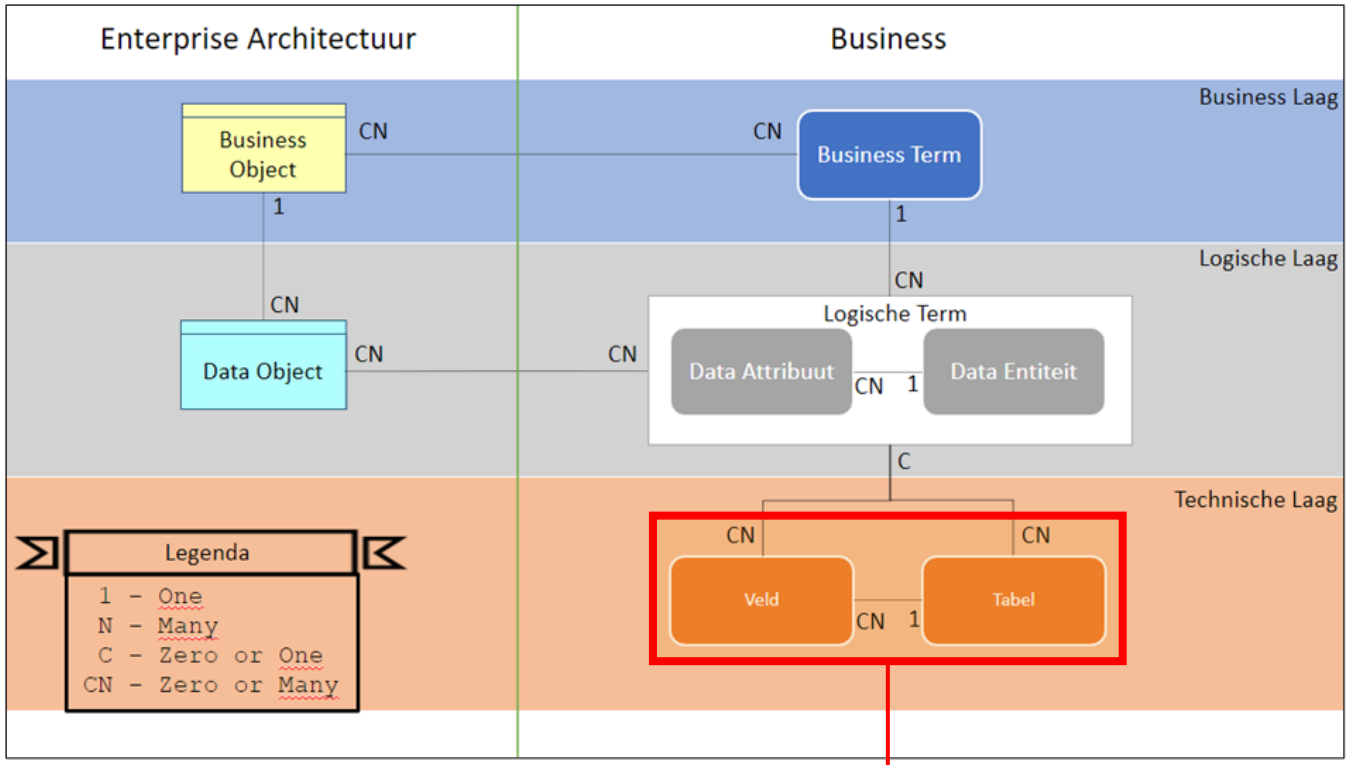
Three layers: business, logical and technical

The PostNL Data Catalog is composed of three different layers:

- **Business layer**
  The terms commonly used by the business and in everyday usage.

- **Logical Layer:**
  The necessary terms (entities and attributes) to link the business terms to the technical layer.

- **Technical layer**
  The fields and tables as they actually appear in the systems and applications.

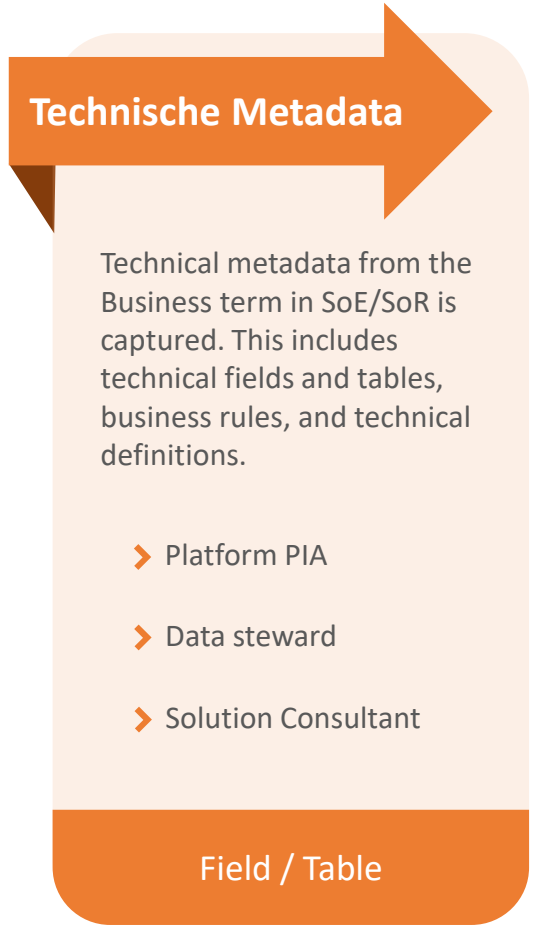


The focus in this document is specifically on the 'Technical layer.' In other words, what is a technical term and what is the associated metadata?

Technical terms

Basis principles (1/2)

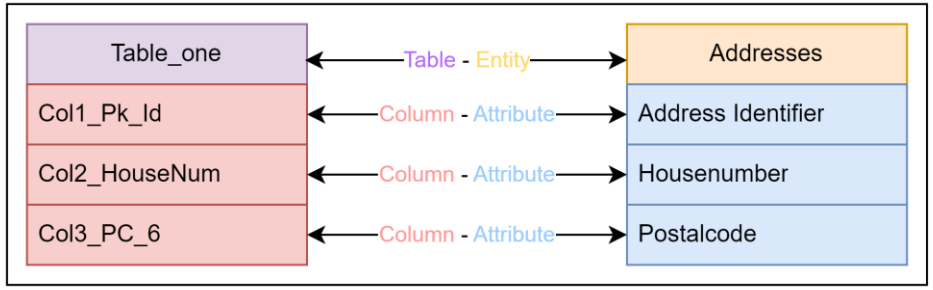
- The technical layer is loaded via the datalake. This concerns the tables and fields that can be found on the physical data layer.
- Despite the data being loaded through the datalake, it is a requirement that the structure as it stands in the datalake matches how it exists in the System of Record. This way, the connection between the Data Catalog and the SoR is simulated. This also means that no tables and fields should be present that have been added to the Datalake afterwards.
- The population of the technical layer (terms + metadata) is largely established during technical loading, inheritance, or bulk mutations. Nevertheless, a portion of it will need to be done manually (e.g., specifying the correct privacy classifications).
- A technical term can have two forms: in the form of a table or in the form of a column.
- Together, the technical terms constitute the physical data found in a system or application. The physical data model describes how data is stored in the database behind the application. Often, this can also indicate how the tables are linked to each other or what data type is expected in the columns.



Technical terms

Basis principles (2/2)

- Technical terms are given system names that are often (too) technical for 'ordinary' readers. To simplify this, they are linked to the logical terms of the logical layer.



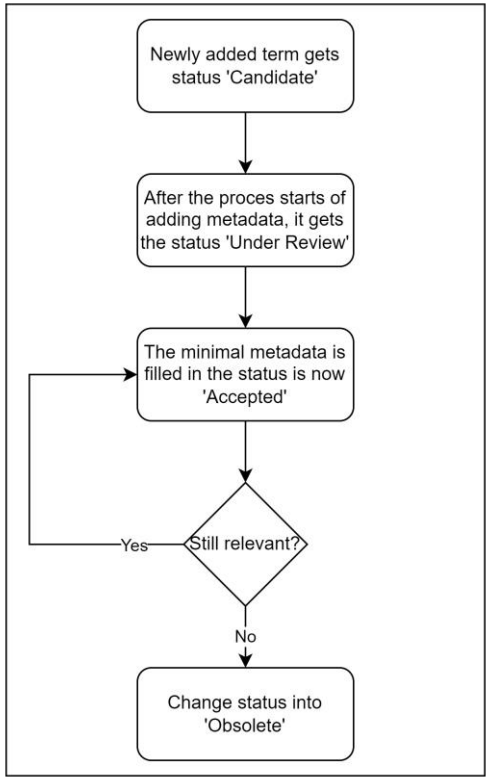
Sources can be loaded or synchronized into the Data Catalog manually or according to a fixed schedule. When these sources are synchronized again, the already filled metadata is not overwritten.

Technical terms

Status

Een term kan in verschillende fases verkeren. Met de status wordt aangegeven of het bijvoorbeeld een geaccepteerde term met metadata is of dat het een term is die '*Under Review*' is.

Status	Definition
Candidate	Initial status of a term. This means that the term has been created or imported. At this stage, there is no examination of any metadata.
Under Review	Stakeholders review the Asset. This means that metadata has been added, but not all required fields have been filled in or are accurate.
Accepted	The term and the mandatory metadata are fully and correctly filled in. The meaning is endorsed by the key stakeholders.
Obsolete	The term is outdated. The metadata for this term remains available for reference. Periodically, a review will determine if cleanup is necessary.
Invalid	Out of scope.
In Progress	Out of scope.
Approval Pending	Out of scope.



Captured metadata

Afgestemd met de Data Management Organisaties

1	**Definition**
2	Explanation
3	Example
4	**Data Domain**
5	**Platform**
6	Platform Supplier
7	**Cloud Service**
8	Cloud Service Supplier
9	**System Of Record**
10	**System of Entry**
11	**Data Labelling**
12	**Personally Identifiable Information**
13	PII Type
14	CIA Classification
15	Confidentiality
16	Integrity
17	Availability
18	Business Ruling
19	Database Type
20	Datatype
21	Validation
22	Language
23	Standard Value Format
24	Date Created
25	Date Changed
26	Made/Changed by
27	Status
28	Source (origin metadata)
29	***Rel Logical Layer – Data Entity***
30	***Rel Logical Layer – Data Attribute***

In consultation with the Data Management Organizations within PostNL, it has been determined that the so-called metadata fields mentioned here on the left should be recorded in the PostNL Data Catalog.

- The red-colored fields are mandatory.
- The black-colored fields are optional.
- The blue-colored fields indicate the relationship with the logical layer.
- Fields 15 to 17 are currently out of scope.
- Fields 24 to 26 are automatically generated by Collibra itself (standard functionality).

The following slides provide further details on each of these fields

Captured metadata

Explanation of the fields (1/6)

Metadata	Description	Guideline	Mandatory
Definition	A clear description of the term in one to two sentences.	• Text field • Definition does NOT contain the term itself • Follows the most common spelling and punctuation • Unique to the context in which the term is described	Mandatory
Explanation	An addition to the definition that clearly describes the term.	•Text field •Explanation does NOT contain the term itself •Follows the most common spelling and punctuation •Unique to the context in which the term is described	Optional
Example	Examples of the term.	•Text field •Follows the most common spelling and punctuation •Unique to the context in which the term is described	Optional
Data Domain	Responsible domain for this specific term.	• Selection list containing the specific Data Domain responsible for this specific term. • Actual overview • If one and the same field can contain a value that may belong to multiple data domains (e.g., the email address of a consumer and the email address of a business partner), the data domain is filled in based on the privacy perspective that is considered the 'most risky.' In this example, the field 'email address' would be assigned the data domain 'Commerce - Customer (consumer).	Mandatory

Captured metadata

Explanation of the fields (2/6)

Metadata	Description	Guideline	Mandatory
Provided by Platform	Indicates on which platform the system is developed.	• Selection list containing all systems and applications used within PostNL on which development can be defined.	Mandatory
Provided by Supplier	Indicates who the developer of the platform is.	• Selection list containing all software suppliers with whom PostNL collaborates or has collaborated.	Optional
Hosted on Platform	Indicates in which cloud environment the system is hosted.	• Selection list containing all systems and applications used within PostNL for hosting.	Mandatory
Provided by Cloud service supplier	Indicates who the supplier of the cloud environment is.	• Selection list containing all cloud solutions with which PostNL collaborates or has collaborated,	Optional
System of Record	Application where the truth is captured and from which distribution to receiving systems and applications takes place.	• Selection list containing all systems and applications used within PostNL defined as Golden Records. • Can contain the same value as the System of Entry (SoE). • If the term has a lineage up to the technical layer, this field is mandatory.	Mandatory
System of Entry	Application where the data is initially entered or generated.	• Selection list containing all systems and applications used within PostNL for data entry. • Can contain the same value as the System of Record (SoR). • If the term has a lineage up to the technical layer, this field is mandatory.	Mandatory

Captured metadata

Explanation of the fields (3/6)

Metadata	Description	Guideline	Mandatory

Data Labelling	Indicates the label applicable to data protection.	• Selection list with the values:
		  • **Public*:** Corresponds to data with ‘none’ confidentiality value
		  • **Internal**: Corresponds to data with ‘low’ confidentiality value
		  • **Confidential****: Corresponds to data with ‘medium’ confidentiality value
		  • **Secret**: Corresponds to data with ‘high’ confidentiality value
		• Must be aligned with and comply with the rules and principles of the PostNL Cybersecurity Office.
		• ** If the column contains PII data, it should be labelled with minimum "Confidential".
			Mandatory

Personally Identifiable Information	Indicates whether it contains personal data.	• Selection list with the values:
		  • Yes, does contain Personally Identifiable Information.
		  • No, does not contain Personally Identifiable Information.
		  • Uncertain if it contains Personally Identifiable Information.
		  • Unknown.
			Mandatory

PII Type	Indicates the type of personal data it contains.	• Multiple-choice selection list with the values:
		  • Business Partner
		  • Consumer
		  • Employee
		• Can be expanded further in the future.
		• * If 'Personal Identifiable Information' is filled in, this field is mandatory.
			* Mandatory

Captured metadata

Explanation of the fields (4/6)

Metadata	Description	Guideline	Mandatory
CIA Classification	The classification of data and related systems is determined by the impact on PostNL when the requirements for Confidentiality, Integrity, and Availability are not met.	• Selection list with the values Baseline and Above Baseline. • When the impact of any of the three categories (Confidentiality, Integrity, or Availability) is high, the CIA Classification is Above Baseline. In all other cases, Baseline is sufficient. • Must be aligned with and comply with the rules and principles of the PostNL Security Office.	Optional

Confidentiality	What is the impact when there is unauthorized disclosure of information.	• Selection list with the values: low, medium, and high:
		  • **None:** The loss of Confidentiality, Integrity, or Availability is expected to have no impact
		  • **Low:** The loss of Confidentiality, Integrity, or Availability is expected to have a limited adverse effect on business processes, assets, or individuals.
		  • **Medium**: The loss of Confidentiality, Integrity, or Availability is expected to have an adverse effect on business processes, assets, or individuals.
		  • **High**: The loss of Confidentiality, Integrity, or Availability is expected to have a severe or catastrophic adverse effect on business processes, assets, or individuals..
		• Must be aligned with and comply with the rules and principles of the PostNL Security Office.
			Optional

Integrity	What is the impact when there is unauthorized alteration or destruction of information.		Optional
Availability	What is the impact when there is a disruption in access for the use of information or information systems.		Optional

Captured metadata

Explanation of the fields (4/6)

Metadata	Description	Guideline	Mandatory
CIA Classification	The classification of data and related systems is determined by the impact on PostNL when the requirements for Confidentiality, Integrity, and Availability are not met.	• Selection list with the values Baseline and Above Baseline. • When the impact of any of the three categories (Confidentiality, Integrity, or Availability) is high, the CIA Classification is Above Baseline. In all other cases, Baseline is sufficient. • Must be aligned with and comply with the rules and principles of the PostNL Security Office.	Optional

Confidentiality	What is the impact when there is unauthorized disclosure of information.	• Selection list with the values: low, medium, and high:
		  • **Low:** The loss of Confidentiality, Integrity, or Availability is expected to have a limited adverse effect on business processes, assets, or individuals.
		  • **Medium**: The loss of Confidentiality, Integrity, or Availability is expected to have an adverse effect on business processes, assets, or individuals.
		  • **High**: The loss of Confidentiality, Integrity, or Availability is expected to have a severe or catastrophic adverse effect on business processes, assets, or individuals..
		• Must be aligned with and comply with the rules and principles of the PostNL Security Office.
			Optional

Integrity	What is the impact when there is unauthorized alteration or destruction of information.		Optional
Availability	What is the impact when there is a disruption in access for the use of information or information systems.		Optional

Captured metadata

Explanation of the fields (5/6)

Metadata	Description	Guideline	Mandatory
Business Ruling	The business rules that apply to a term.	• Text field • Example: Dutch Postal Code: Consists of four digits and two letters. The first digit cannot start with 0. The letter combinations 'SS,' 'SD,' and 'SA' are not used.	Optional
Standard Value Format	The standard format that applies to a term.	• Text field • Specifies the format in which the term should be filled out. • Can be filled in as a regular expression. • Example: Dutch Postal Code: /^[1-9][0-9][0-9][0-9][A-Z][A-Z]$/gm	Optional
Database type	Indicates the type of structural format the database has.	• Selection list with the values: Network Database, Object-oriented Database, or Relational Database.	
Data type	Indicates the type or kind of data allowed in a field.	• Indicates the type or kind of data allowed in a field.	

Captured metadata

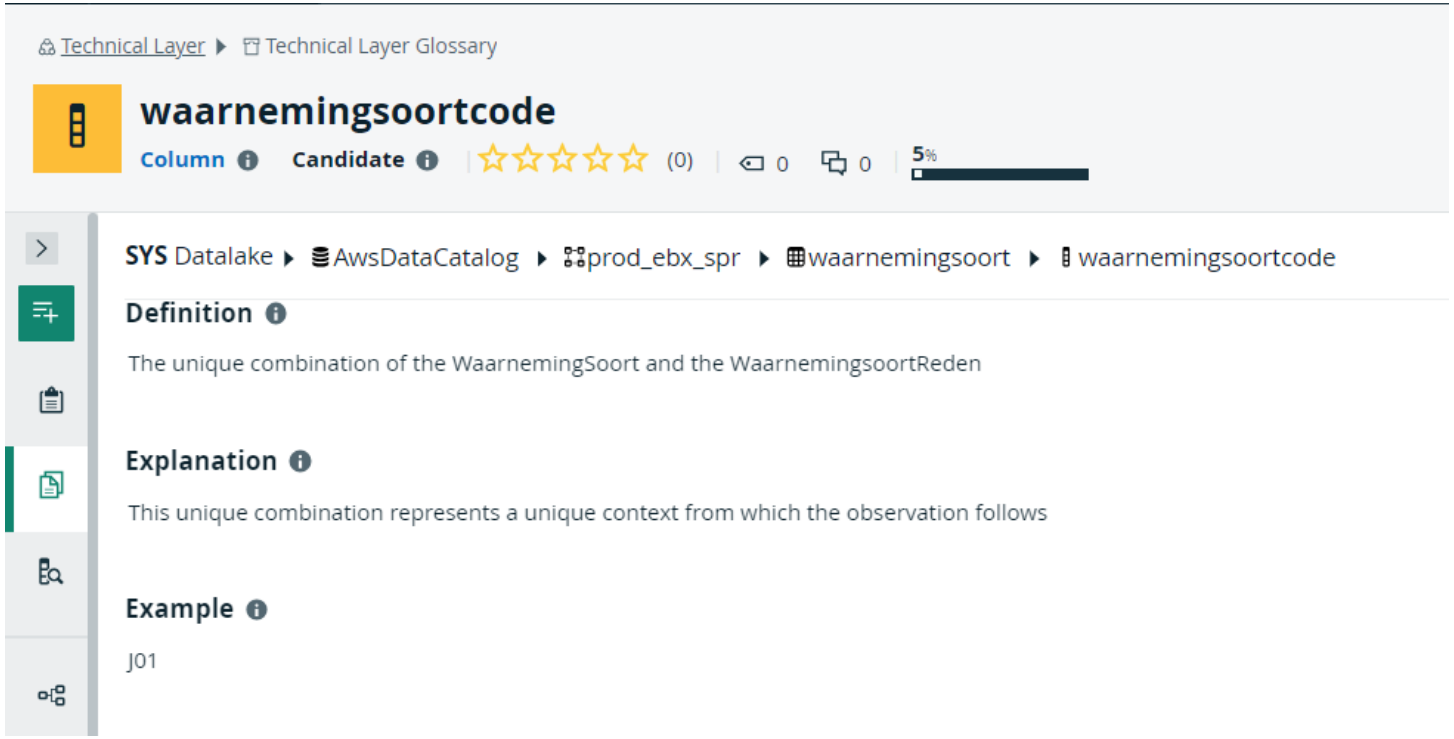
Explanation of the fields (6/6)

Metadata	Description	Guideline	Mandatory
Synonym	The relationship with terms that have a similar or identical definition within the same Data Domain.	• Has a relationship with another term in Collibra. This requires that the term with which a relationship is to be established exists in Collibra. • Can only exist with other terms on the same layer. • The relationship must be indicated for both term A and term B.	Optional
Validation	Indicates whether a field in the source is validated against the 'truth'.	• Boolean with the options yes or no.	Optioneel
Source	Indicates what or who (function/role) is the source of the filled metadata.	• Text field. • Can contain one or multiple sources. • Can refer to functions, roles, and/or systems.	Optioneel
Relations with other layers	The relationships to other terms will be explained on separate slides.	• Has a relationship with another term in Collibra. This requires that the term with which a relationship is to be established exists in Collibra.	Optional

Captured metadata

An example

- A few examples of filled metadata for the technical term (column) 'observation type code'.



Relations

Relation with logical terms

- It is not possible to directly link a Business term to a Technical term (tables and fields) (or vice versa).
  - When there is a need to make the technical implementation at the application and system level (technical layer) transparent, a relationship must be created at the logical layer.
  - Example: The technical implementation of a process location.
- A technical term can be linked to a logical term (data entity or a data attribute) at the logical layer.
  - Data Entity is the logical counterpart of a table.
  - Data Attribute is the logical counterpart of a field.
- Only when a system is loaded at the technical layer can it be linked to the related logical term in the data catalog.

Captured metadata

Explanation of the fields (3/6)

Metadata	Description	Guideline	Mandatory

Data Classification	Indicates the classification applicable to data protection.	• Selection list with the values: low, medium, and high.
		  • **None:**
		  • **Low**: Corresponds to data for internal use.
		  • **Medium**: Corresponds to confidential data.
		  • **High**: Corresponds to secret data.
		• When data is publicly available, it receives either no classification or the low classification.
		• Must be aligned with and comply with the rules and principles of the PostNL Cybersecurity Office.
			Mandatory

Personally Identifiable Information	Indicates whether it contains personal data.	• Selection list with the values:
		  • Yes, does contain Personally Identifiable Information.
		  • No, does not contain Personally Identifiable Information.
		  • Uncertain if it contains Personally Identifiable Information.
		  • Unknown.
			Mandatory

PII Type	Indicates the type of personal data it contains.	• Multiple-choice selection list with the values:
		  • Business Partner
		  • Consumer
		  • Employee
		• Can be expanded further in the future.
		• If 'Personal Identifiable Information' is filled in, this field is mandatory.
			Mandatory

Captured metadata

OUD

Metadata	Description	Guideline	Mandatory

Privacy Classification	Indicates the classification applicable to data protection.	• Selection list with the values: low, medium, and high.
		  • **Low**: Corresponds to data for internal use.
		  • **Medium**: Corresponds to confidential data.
		  • **High**: Corresponds to secret data.
		• When data is publicly available, it receives either no classification or the low classification.
		• Must be aligned with and comply with the rules and principles of the PostNL Cybersecurity Office.
			Mandatory

NIEUW 1.0

Metadata	Description	Guideline	Mandatory

Data Labelling	Indicates the label applicable to data protection.	• Selection list with the values: low, medium, and high.
		  • **Public*:** Corresponds to data with ‘none’ confidentiality value
		  • **Internal**: Corresponds to data with ‘low’ confidentiality value
		  • **Confidential****: Corresponds to data with ‘medium’ confidentiality value
		  • **Secret**: Corresponds to data with ‘high’ confidentiality value
		• Must be aligned with and comply with the rules and principles of the PostNL Cybersecurity Office.
		• * All data within PostNL should be labelled with minimum "Internal"
		• ** If the column contains PII data, it should be labelled with minimum "Confidential".
			Mandatory