SET Academy

Project Proposal - Final Project

Batch: November 2021

Date: 24th of January - 18th of February 2022

Project Title: Smart Data Catalog Enhancement

Propulsion Academy Contact:

Project Type: 3.5 Week Bootcamp Project

Company: sidecar-data.ch

Contact Person(s): Raphael Colsenet - raphael@sidecar-data.ch

Project Description / Business case:

Introduction - Sidecar at a glance

Sidecar is what we call a data catalog. The main purpose of a data catalog is to store valuable metadata about data assets available in a company, in order to help data consumers (data analysts, data scientists, chief data officer, etc.) find the data they need for their analysis. Having a clear, accurate and up-to-date view on the assets available can save hundreds and hundreds of wasted hours every month.
Sidecar data catalog has 4 big features :

- Build a comprehensive view of the data assets across the company
- Enrich data assets with meaningful business context
- Allow users to explore a unified view of all data assets
- Discover and manage sensitive data to protect data assets

Technical metadata are collected automatically but they are not very valuable. To leverage the value of a data catalog, the business context of a data point is crucial to the users. These metadata are usually manually documented in the catalog by a Data Steward (person in charge of the data sets with business knowledge about the data).

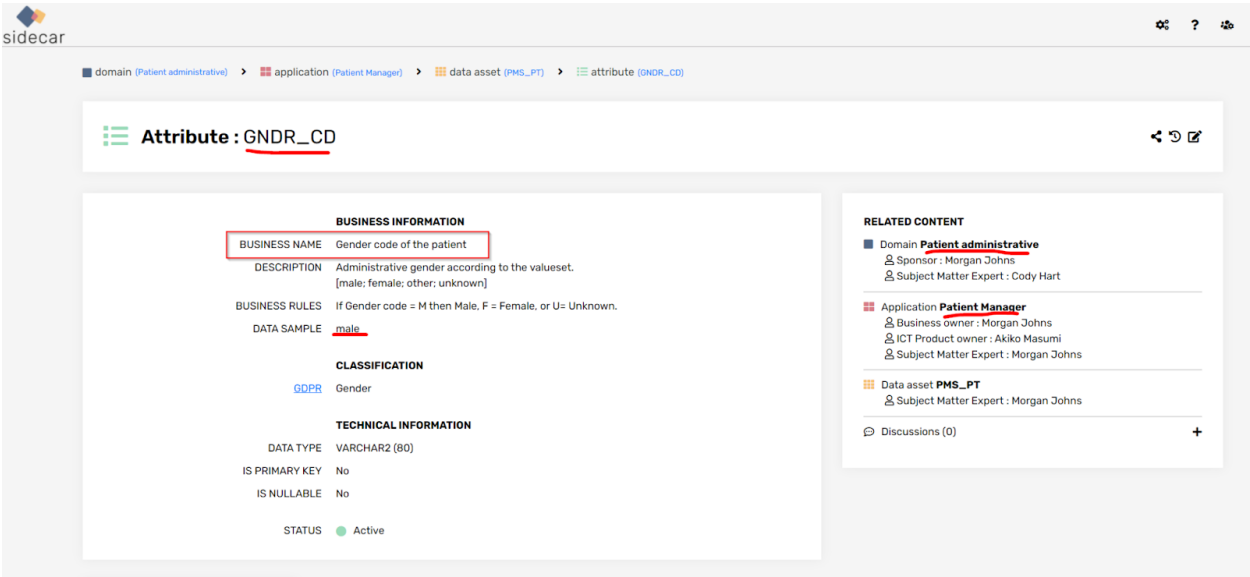
Finding a way to help the data steward to describe the data would be very precious and a real accelerator!

Project Goals:

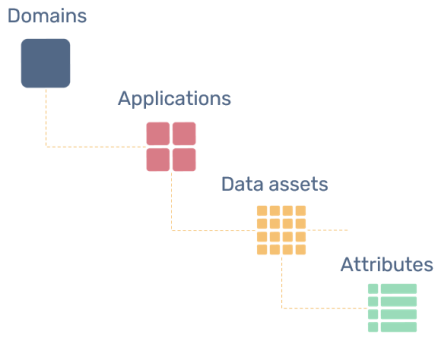
The goal of this project is to explore how machine learning can help the data steward fill the business metadata of a data point by suggesting a potential business name.

Let's take a simple example:
I have a database column named "GNDR_CD", this column is part of a table called PMT_PT, and this table is used in the application "Patient Manager". Based on this information, as well as perhaps some external sources and the data itself, the idea is to generate a business name for the column "GNDR_CD" such as "Gender code of the patient".
The data steward will then have to validate or adjust the definition before saving it in Sidecar.

Sidecar view of the column "GNDR_CD"



The data in the catalog is organized hierarchically, the most granular level is the attribute (column, cell, …). The attributes are grouped into Data assets (table,file,..). The data assets are grouped into Applications and then into Domains.



Potential data that can be used to define the business name:

- The technical metadata of the attributes or the data asset (technical name, e.g.: “GNDR_CD”)
- The data stored in the attributes or the data assets (e.g.: “male”)
- Name, description, etc.. of the Application and Domain (e.g.: “Patient Manager”, “Patient administrative”)
- External datasets (e.g.: list of synonyms, list of iso codes,...)
- Patterns defined to help identify the business name (e.g.:credit card pattern, telephone number, Iban,...)
- Other ideas are welcome, be creative!

Milestones:

- Define an algorithm that can be used in Sidecar to automatically generate the business name of the attributes or the data assets.
- Produce new metadata from the algorithm to help the data steward decide if the business name is relevant or not (e.g.: matching score, ranking, etc…)

Constraints:

- Use technologies that can be easily encapsulated and deployed in Sidecar