# Use of Knowledge Graph in Data Governance

A picture containing small, sitting, grass, clock

Description automatically generatedThe concept of Knowledge Graph exists for long time. In recent time Google has made a significant paradigm shift in their search engine by introducing similar perception. In a nutshell, knowledge graph is a collecting information about different entities and their relationship to one another. Entities can be book, person, movies, landmarks, celebrities, cities, sports teams, buildings etc. This graph allows you to store information in a graph model and use it easily to navigate through interconnected entities. You can check the blog “Introducing the Knowledge Graph: things, not strings” (<https://www.blog.google/products/search/introducing-knowledge-graph-things-not/>) to learn more about Google’s implementation.

This post talks about knowledge graph in the context of information governance. Here the phrase “knowledge graph” used in a more general sense, as a technology and services within IBM Cloud Private for Data (CPD).

The explosion of data and data analytics are making data governance more important, and more difficult. Most organization holds unprecedented amounts of data sets scattered across different sources and wide range of users consuming these data. The size and complexity of these data sets makes old style data dictionaries and Entity Relationship Diagrams (ERD) out dated. The information (data) governance within the self-service eco system of CPD provides overall management of data availability, relevancy, usability, integrity and security in an enterprise. It helps enterprise manage their information knowledge, and answer questions on “What, How, Who, When, Why” (Five Ws) of basis of information- gathering and problem-solving.

Data governance practices provide a holistic approach to managing, improving and leveraging information to help an enterprise's overall data management efficiency. The information governance policies extend across all kind of business assets. It’s not only governed data to ensure compliance; governance policies extend to enterprise information asserts like ML models, Notebooks, R Shiny Application etc., which is an enabler of batter business outcomes.

Now let’s start with building a knowledge graph based on the data governance. In this journey you will use some data set from a remote Db2 data source that captured from mortgage applications processing. This data set and used case are similar one that available in IBM Information Center: <https://www.ibm.com/support/knowledgecenter/en/SSQNUZ_2.1.0/com.ibm.icpdata.doc/zen/tutorials/mortgage/mortgage.html>. Goal here is to build a knowledge graph to explore all mortgage related assets based on the data catalog information within CPD. For the sake of this post and simplicity, you will use following two tables only:

|  |  |
| --- | --- |
| MORTGAGE\_CUSTOMER | MORTGAGE\_DEFAULT |
|  |  |

Figure 1: Table structure

There are two distinct phases for enabling knowledge graph.

1. Collect and Organize Data – In this phase you wear many hats to prepare the data for knowledge graph.
   1. Collect Data – The mortgage data is residing on a remote Db2 data source. As a data engineer you need to create the connection to the Db2.
   2. Discover Assets – The data discovery service helps catalog metadata across the enterprise to search and govern the data. Run data discovery on the newly created Db2 connection, which will pull metadata from Db2 database and populate the CPD catalog.
   3. Build Business Glossary – The business glossary is nothing but a catalog of assets that defines the characteristics of an enterprise. As a data stewards you create terms and categories to form a logical structure between actual table definition and business terms. For instance, the ‘ID’ columns in table definitions. What it’s actually describing? Unless you create a term with proper description, this column name does not make any meaningful impact during catalog search or data analysis. Creation of proper categories and terms are the deciding factor for a successful knowledge graph building. Categories provide a ways to browse and understand the relationships among business assets in the business glossary. Whereas terms describe characteristics of the enterprise and fundamental building blocks of glossary. You will spend lots of time to create a hierarchy that could reflects how a user might search for information. A common strategy is to divide the business by subject areas, such as mortgage\_customer and mortgage\_default etc.

First create some categories, to give appropriate meaning to mortgage related data.

Create the category **Mortgage\_category** for the mortgage data.

1. Go to **Organize** > **Business glossary** > **Categories**
2. Click **Create Category**
3. Provide name, and optionally parent category and description as below:

A screenshot of a cell phone

Description automatically generated

1. Click **Save**

Create sub categories to describe each tables under mortgage data set. In this case those are MORTGAGE\_CUSTOMER and MOSTGAGE\_DEFAULT.

Start with create a sub category for MORTAGE\_CUSTOMER.

1. Go to **Organize** > **Business glossary** > **Categories**
2. Click **Create Category**
3. Provide name, and optionally parent category and description as below:

A screenshot of a cell phone

Description automatically generated

1. Click **Save**

Similar as above, crate another sub category for MORTGAGE\_DEFAULT.

Next, create terms for all columns, starts with column name ID.

1. Go to **Organize** > **Business glossary** > **Terms**
2. Click **Create Term**
3. Provide name, select the parent category, status, and optionally add description

A screenshot of a cell phone

Description automatically generated

1. Click **Save**

Similar way creates terms for each individual columns in all tables.

* 1. Auto Assign Terms - Re-run the discover asset process, it will automatically associate assets to business term that created previously. Initially these associations are represented as candidates, which stewards can review and adjust if necessary. Terms are represented as candidates if their confidence level matches or exceeds 40%. Candidates are automatically assigned if their confidence level matches or exceeds 80%.
  2. Publish Assets – The results of automatic term assignment are not visible to all users unless those published. After the automated discovery process completes, candidate terms are suggested, approved, and published those assets will be available in the enterprise search.

1. Self-service Access of Data – Now imagine you are a data scientist who needs to create a model to predict mortgage default rate. In this scenario enterprise search should be your first place for figure out different mortgage related data assets available in the catalog.

2.1) Start with search for word mortgage in enterprise search. This will display all the data assets that goes with mortgage.

2.2) Look for category Host in the search results. Host will be the root of the mortgage data set, which point to a JDBC connection to the data source.

2.3) Click on the Relationship Graph for the host, that will begin your journey to knowledge graph.

2.4) Click on the Explore relationship icon (on top right corner) to see the relationship graph for host category. You should click on each + sign to expand the relationship between mortgage related assets. Assets could be host, database, database schema, tables, columns, categories, terms etc. Click on the individual asset box to get detail information about the respective asset. As you move through the relationship graph, system will try to find out identifier between tables. Each assets in the relationship graph will give you enough idea to decide which table, columns need to consider for the building your mortgage prediction model.

Figure 2.1 is an example of relationship graph that generated based on mortgage related information available in the catalog.

A screenshot of a cell phone

Description automatically generated

Hope the above information give you some idea on create relationship graph, which is indeed similar to the knowledge graph. Remember, the success or failure of the relationship graph is depending on the how good is your business glossary.