Knowledge Graphs for Social Good

Building A Simple Knowledge Graph with UN Data

Quick Start Example, Common Methodologies, and Tooling

Colin Puri (colin.puri@gmail.com, colin.puri@accenture.com)
Accenture Labs, R&D Principal
This presentation serves to empower users in demonstrating an approach on how to get started working with the UN datasets for the process of scoping and modelling data into a knowledge graph. This walkthrough tutorial will focus on how to break down key questions and data into concepts, relationships, and the properties of a knowledge graph with respect to the Sustainable Development Goals (specifically poverty measurements) and linking to news articles.

Knowledge graphs

Use cases
- Representation of highly related data
- Filtering
- Faceting
- Searching
- Inferencing

Provides
- Increased precision and accuracy
- Accessibility with API integration
- Democratization of knowledge

Process of Applying Knowledge Graphs

Creation:
- Model schema
- Define entities and properties from concepts
- Define relationships for inferencing

Hydration:
- Source, process, and stage data
- Publish ingestion services which consume
- Populate the instance graph

Application:
- Query instance graph
- Allow for data insertion (public access)
- Create and Manage API's
Overview

• Introduction
  • Knowledge Graph Enablement

• Example with UN Data –
  • Walkthrough Scenario
  • Scope
  • Define
  • Hydrate & Apply

• Wrap Up
  • Example tooling
  • Conclusion
Knowledge Enablement: Why do we need a knowledge graph

Knowledge Graph Core Strengths

Data
- Represents highly related data
- Allows for many interconnected inferences

Searching
- Provides filtered information
- Directs searches to key concepts
- Provides query expansion and enrichment in a domain model
- Provides significant increased precision with information retrieval

Steps of Defining and Working with a Knowledge Graph

1. **Scope**
   a. Define problem statement and questions
   b. Breakdown questions into requirements and define search needs
   c. Identify key elements and domain concepts
   d. Connect requirements to technical requirements

2. **Define & Model**
   a. Relate elements and domain concepts to available data and data schema
   b. Model domain with data models
   c. Create entities (e.g. nodes)
   d. Define entity properties from data and business concepts
   e. Create relationships between entities/nodes (as key facets to filter across – subject, predicate, object type relationships)

3. **Ingest & Instantiate**
   a. Define data transformations necessary and model ETL process
   b. Load data using data mapping models

* Define & Model (cont.)
Example with UN Data
Challenge Scenario: Learn Your Data Context

User Scenario Summary:
Imagine that your team has recorded and compiled data from the field on poverty and you would like to learn more about the indicators and why the numbers are what they are. Specifically, what countries are recording metrics for poverty, is it broken down by gender, and is there any external information that can help provide an explanation or basis for the numbers you are seeking.

Today you would ask colleagues, attend seminars, etc. but this is time consuming and there isn’t enough time in the world to talk to everyone to get a complete picture. You could read internet but finding documents can be difficult at times and time consuming. What you really want is to quickly receive a summary broken down by country and relating news articles.

Key Scenario Questions:
How do you model and search poverty measurements in the UN Data?
Is there a way to understand equality within the poverty data?
How do you find, summarize, or attach related news articles?
Scope
Step 1: Scope

Human modelling with a symbolic approach that requires inferencing means that a Knowledge Graph is the best approach.

How to model a problem space into a Knowledge Graph

Definition Scope

Define the scope of the challenges space by following these steps:

a. Gather the key questions
b. Breakdown questions into key requirements and define needs

c. Identify key elements and domain concepts
d. Analyze data sources to understand if data contains key elements and business domain concepts
e. Proceed to modelling, revisiting previous steps as needed*

*The key to a successful modelling of a schema is to start out small and rapidly iterate between scoping, modelling, and testing queries.
### Step 1: Scope – Breaking Down the Top-Level Questions

**Gather the Key Questions, Identify Key Business Elements & Concepts**

**A. How do you model and search poverty measurements in the UN Data, per country –**

**Summary:** There is a desire to search based on the poverty data versus other versus, perhaps poverty should be defined as an entity since Target 1.1 models poverty and also have attribute values that distinguish them from other data

<table>
<thead>
<tr>
<th>Key Requirement</th>
<th>Key Elements &amp; Concepts</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search for the poverty data vs. other country data</td>
<td>Model “Measurement Indicator” and “Country” as a concepts</td>
<td>Model a “Targets” and “Goals” as concept in the graph schema as a reusable component since there are also many other targets and goals as apart of the SDGs. Additionally, many measurement indicators can be referred to in many ways, therefore model a topic or synonym as a entity with a relationship between that concept and the policy concept</td>
</tr>
</tbody>
</table>

**B. Is there a way to understand equality within the poverty data as it changes –**

**Summary:** There is an additional request to understand the data for inequality. Given that the data is broken into gender. The data for Target 1.1 of the SDGs breaks data down according to the Goal, Target, Indicator, Measurement Value, time frame, and for which gender the data refers

<table>
<thead>
<tr>
<th>Key Requirement</th>
<th>Key Elements &amp; Concepts</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter and infer data according to gender</td>
<td>Model relationships such that a country “has a Measurement”. Also there should be properties for Male or Female or Both for any given measurement</td>
<td>Model properties to the concepts and relationships with valid time frames to what information can be filtered through time</td>
</tr>
</tbody>
</table>
Gather the Key Questions, Identify Key Business Elements & Concepts

C. How do I understand the key measurements and How do you to find, summarize, or attach related news articles for comprehension -

Summary: This implies that there overlap with news articles and the key measurements for poverty.

### Key Requirement
Infer that one or more news articles have relevant information pertaining to an indicator

### Key Elements & Concepts
Model an “Attribute” concept for keywords that describe a goal, target, or measurement indicator. Include relationships between “Attribute” and “Measurement Indicator” to create a contextual web of related concepts

Model “Article” to hold information about news articles and include relationships to attributes that describe that article

Model “Synonym” relationships from concepts to the “Attribute” concept that may have multiple ways of referring to them

### Extension
Model a hierarchy of concepts in the final graph so you can relate synonyms for added inferencing capabilities.
Define & Model
Step 2: Define and Model

Once the core concepts are determined that should be modelled based on the set of key high-level questions, relate the desired concepts to the available data, data format, and data schema to determine viability.

Define Model

Create the schema space by following these steps:

a. Relate elements and domain concepts to available data and data schema
b. Model domain with reusable data models
c. Create entities (e.g. nodes)
d. Define entity properties from data and business concepts
e. Create relationships between entities/nodes (as key facets to filter across – subject, predicate, object type relationships)
f. Proceed to modelling, revisiting previous steps as needed*

*The key to a successful modelling of a schema is to start out small and rapidly iterate between scoping, modelling, and testing queries
### Step 2: Define and Model

Relate elements and domain concepts to available data and data schema

<table>
<thead>
<tr>
<th>SDG Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Goal</strong></td>
</tr>
<tr>
<td><strong>Target</strong></td>
</tr>
<tr>
<td><strong>Measurement Indicator</strong></td>
</tr>
<tr>
<td><strong>Country</strong></td>
</tr>
<tr>
<td><strong>Attribute</strong></td>
</tr>
</tbody>
</table>
Step 2: Define and Model

Relate elements and domain concepts to available data and data schema
When merging two or more ontologies, focus on discovering overlapping entities that can bridge the gap

A. **Attribute**: A one word or phrase description that can be used as synonyms that describe a goal, target, measurement or article

B. **Article**: The concept for holding for a news article that relates its attributes and other concepts
Step 2: Define & Model – Start Small & Keep it Simple

Defining and Modelling is an iterative process by which the Knowledge Engineer defines a schema from an initial set of key questions that have been broken down into key concepts and from looking at source data that will be ingested. The Knowledge Engineer also needs to understand when to model data as an entity or property or relationship based on these key starting rules:

**Entities vs. Properties vs. Relationships**

<table>
<thead>
<tr>
<th>Function</th>
<th>Concept</th>
<th>Entity</th>
<th>Property</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facet &amp; Group</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infer</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

**Filtering** – if a data element is utilized for searching and traversing through concept then it should be modeled as a property

**Faceting & Grouping** – if a data element is used as a high level concept to group things together then it should be modelled as an entity

**Inferencing** – If a relating one or more data elements then it should be a relationship

*Edge directionality is only important for Query generation, not the core storage of knowledge, information, and concepts
Hydrate & Ingest
Step 3: Ingest & Instantiate

Once the core concepts are modelled then the data ingestion modelling may begin by creating data flows, modelling ETL jobs, and publishing APIs. The process for scaling ingestion and instantiation is as follows:

1. **Source data** and documents
2. **Process documents** via a document intelligence process
3. **Stage Data** in a normalized, transportable process
4. **Model the ETL dataflow** process to consume the normalize files
5. **Publish ETL dataflow**
6. **Post data to graph** via published ETL data flow API and post information to the graph

**Document Intelligence Process** → **Extracted Information (Transportable – CSV)** → **Modelled ETL Data Flow** → **Post data to graph**
Step 3: Ingest & Instantiate

Processing of documents further down the dataflow pipeline requires that data formatted for easy consumption.

Example of Incoming Data

- In order to be useful, source data must be converted into a consumable CSV with information aligned into columns according to desired model type for entity or property.
- Data is human readable, but is structured for quick parsing by ETL engine.
- Can be converted using custom scripts or further refined in the document intelligent processes.

Source Documents

Extracted Information (Transportable – CSV)

Create ETL Dataflow and write to graph
Example Working Graph
Wrap Up
Example tooling

Knowledge Engineering

Graph Databases

ETL (Extract, Transform, Load)
Conclusion

Scoping, Defining, & Modelling

Key Questions
- How do you model and search poverty measurements in the UN Data?
- Is there a way to understand equality within the poverty data?
- How do you find, summarize, or attach related news articles?

Graph Schema

<table>
<thead>
<tr>
<th>Source Data</th>
<th>Enterprise Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDG Goal 1, Target 1.1, Indicator 1.1.1</td>
<td>Measurement Indicator, Measurement Instance, Attributes, Country…</td>
</tr>
<tr>
<td>News articles</td>
<td>Articles, Attributes</td>
</tr>
</tbody>
</table>

Ingesting, Populating, & Applying

Extracted Information (Transportable – CSV)

Modelled ETL Data Flow

Looking Ahead (extra points to consider)

A knowledge graph and knowledge management system is for naught if it cannot be accessible, queried, and exposed for consumption. A knowledge graph is only useful in the context of enabling an application (e.g. conversational, exploration, etc.)

Choose solutions and graph schemas that allow
- Allow for a consumer to query data
- Allowing data owners to insert data
- Allow for integration with other applications (e.g. Virtual Assistants) and data consumers

Original Query: What is the poverty for Angola?
Boolean Query: ‘Measurement Synonym_Of Attribute Name Contains “Angola”’...
Questions?

For additional questions or follow up, please email

colin.puri@gmail.com