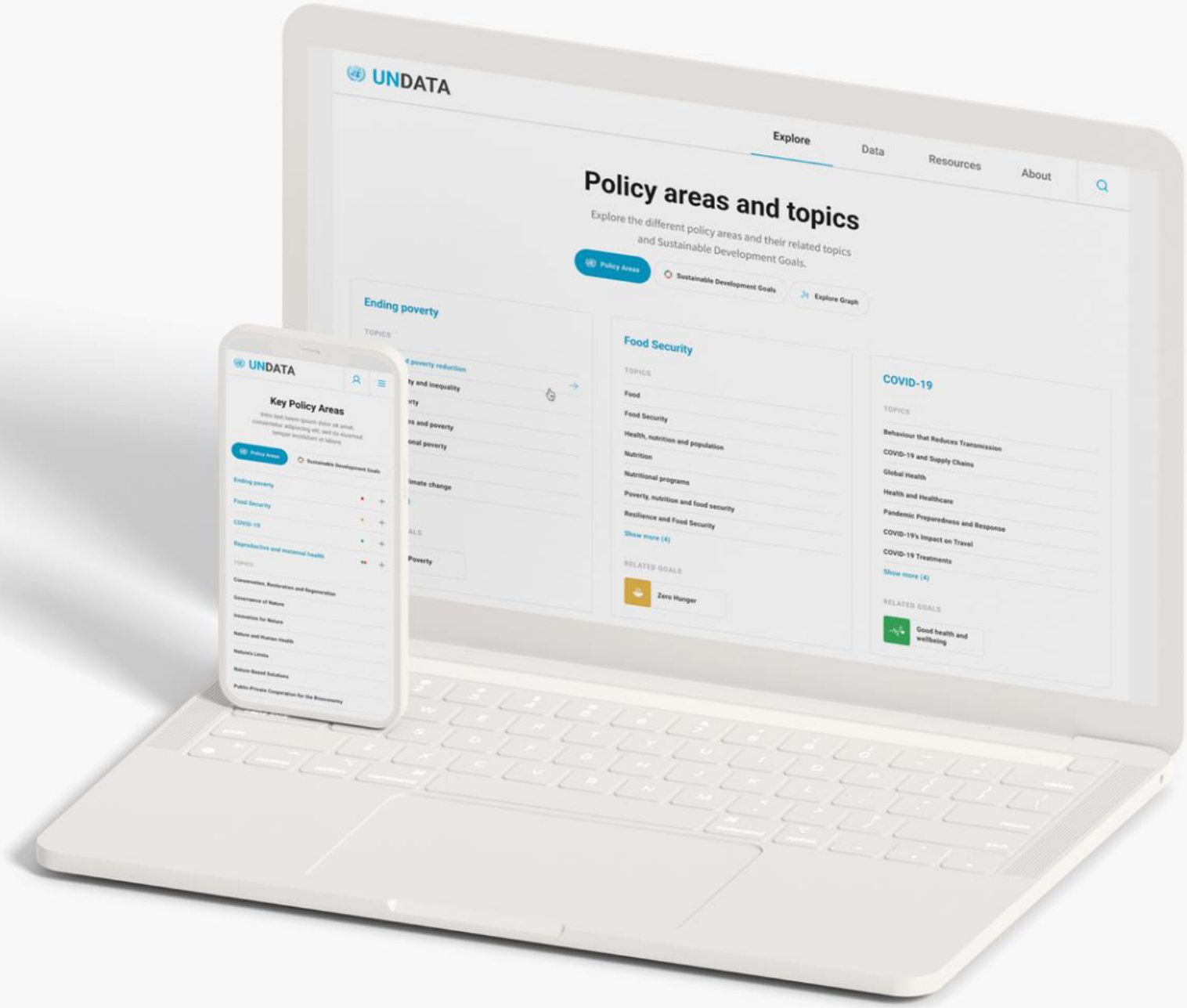
United Nations | Department of Economic and Social Affairs

SDMX meets Data Commons (at UNDATA)



UNDATA

United Nations | Department of Economic and Social Affairs

Context

- Response to UN Secretary-General's Data Strategy
- Informed by the Roadmap for Innovating UN Data and Statistics
- Endorsed by UN Secretary-General's Executive Committee

UNDATA

United Nations | Department of Economic and Social Affairs

Overall objectives

- Increase visibility and access to authoritative sources of statistical data and metadata
- Improve search and analytic capabilities for policy and decision makers
- Enable interoperability of statistical data from across the UN system, Member States and other partner organizations
- Enhance data value through meaningful interlinkages across global, regional and national data portals

UNDATA

United Nations | Department of Economic and Social Affairs

Working principles

- Harness the UN's authority, credibility, and name recognition to unify efforts and collaborate across all stages of the data life cycle
- Ensure trust in data through data quality and adherence to standards
- Develop capacity of all stakeholders to both contribute and use a common, modern data infrastructure

UNDATA

United Nations | Department of Economic and Social Affairs

Building on existing standards, infrastructure and communities of practice

- Collaborate with domain experts in data modeling and integration tasks to validate transformations and mappings
- Use .Stat Core as dimensional data repository, leveraging SDMX standards
- Deploy .Stat Core and associated tools on UN Global Platform infrastructure
- Engagement with SIS-CC community to build capacity to manage data lifecycles within the .Stat core

UNDATA

United Nations | Department of Economic and Social Affairs

Towards a distributed UNdata Knowledge Graph

The objective is to capture the concepts and relationships required to:

- Establish explicit and implicit links to external resources, thus making data more easily findable, searchable, and usable.
- Build applications that efficiently access related data across multiple domains using linked open data techniques
- Generate insights by reasoning over complex relationships.
- Incrementally add new data and evolve the data schema to accommodate new data types and new use cases.

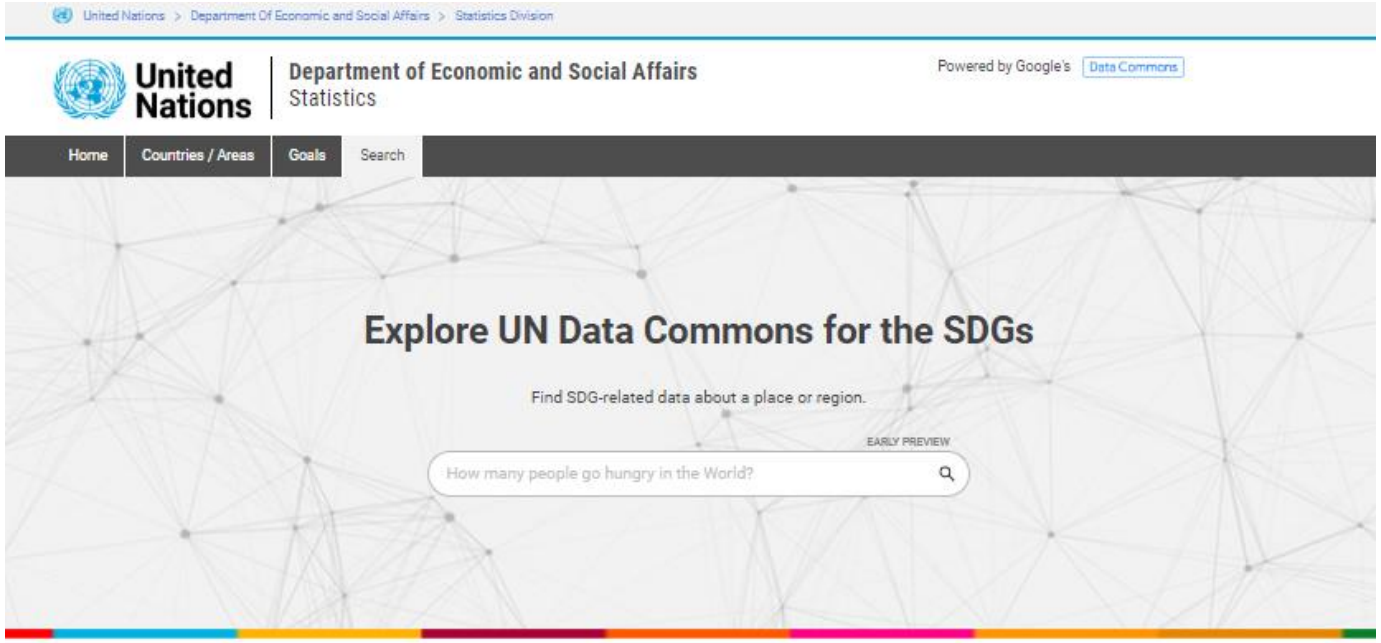
UNDATA

United Nations | Department of Economic and Social Affairs

SDG Summit, 18-19 September 2023

- Launch of a first version of the UN Data Commons, focused on SDG data

  - ✓ Powerful data visualizations
  - ✓ Advanced Natural Language search functionality making it easy for users to access, explore, and utilize the data



https://unstats.un.org/UNSDWebsite/undatacommons/sdgs/

UNDATA

United Nations | Department of Economic and Social Affairs

Next iteration: Towards the 2024 Session of the UN Statistical Commission

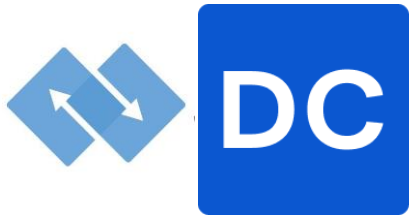
- Fully incorporate dataset from “pilot agencies”, creating new thematic and data provider landing pages
- Provide partner agencies with tools to streamline their data integration workflows
- Enable multi-lingualism

UNDATA

United Nations | Department of Economic and Social Affairs

UN Data Commons data architecture

Underlying framework to organize, integrate, and manage data within the UN Data Commons platform

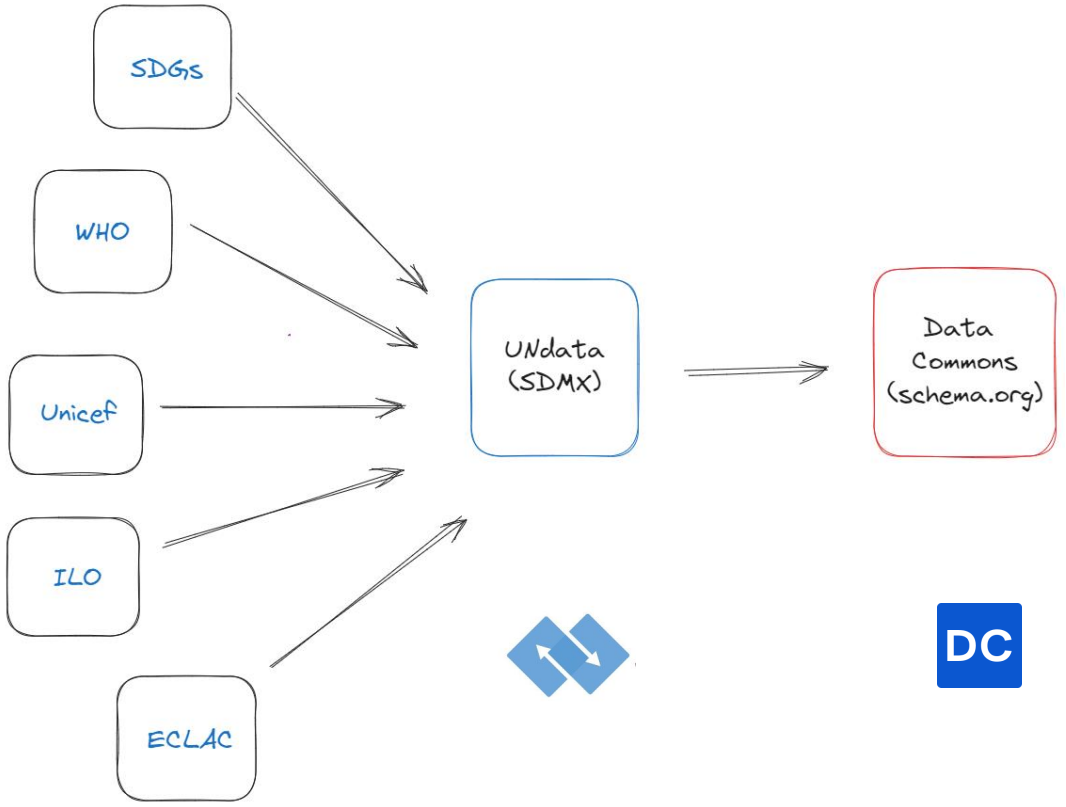


UNDATA

United Nations | Department of Economic and Social Affairs

Objective

Integrate datasets from multiple, heterogeneous sources, by converting them to a common “UNdata” schema, and ingest them into the Data Commons Knowledge Graph (based on schema.org)



UNDATA

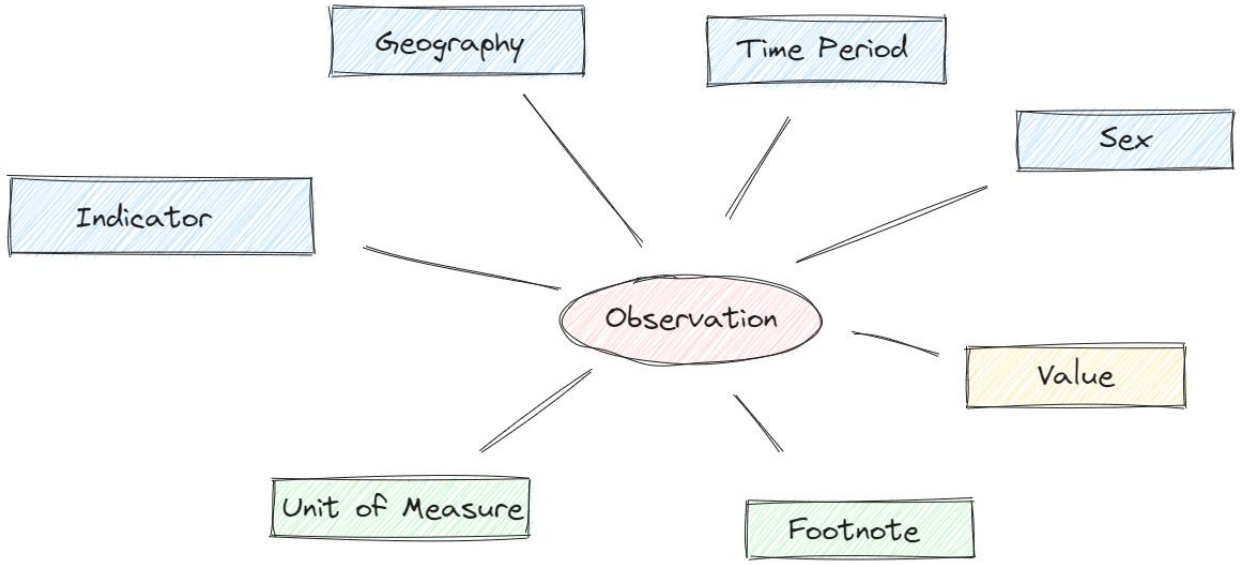
United Nations | Department of Economic and Social Affairs

Statistical Data

- Aggregated
- Dimensional data model
  - Dimensions
  - Attributes
  - Measurements
- Key dimensions (always present):
  - Indicator
  - Geography
  - Time
- Examples of other dimensions: Sex, Age, …
- Key attributes (always present):
  - Unit of measurement,
- Example of attributes: Footnotes…

Indicator	Geography	Time Period	Sex	Value	Unit of Measure	Footnote
Life expectancy	Ghana	2022	Total	…	Percent	lorem ipsum…
Life expectancy	Ghana	2022	Male	…	Percent	lorem ipsum…
Life expectancy	Ghana	2022	Female	…	Percent	lorem ipsum…
Life expectancy	Vietnam	2022	Total	…	Percent	lorem ipsum…
Life expectancy	Vietnam	2022	Male	…	Percent	lorem ipsum…
Life expectancy	Vietnam	2022	Female	…	Percent	lorem ipsum…
Life expectancy	Argentina	2022	Total	…	Percent	lorem ipsum…

Concept schemas



Enumerations

Sex_Code	Sex_Label
_T	Total
F	Female
M	Male

Geography_Code	Geography_Label
G00000020	Afghanistan
G00000050	Albania
G00000060	Algeria
G00000210	Bahamas
G00000380	Brazil

Data integration problem

- A significant part of the work is in aligning terms used to refer to the same or overlapping concepts across different datasets.
- Equivalent concepts (entities) are assigned different identifiers in different databases or vocabularies
- Different communities use different names for the same real-world entities
- Mapping rules are often hidden or not documented at all
  - Describe the correspondences between classification schemes
  - Tracking how classification items have been created, split, merged, or removed from active use

UNDATA

United Nations | Department of Economic and Social Affairs

Mappings between source entities and UNdata entities

Geography_Code	Geography_Label
G00000020	Afghanistan
G00000050	Albania
G00000060	Algeria
G00000210	Bahamas
G00000380	Brazil

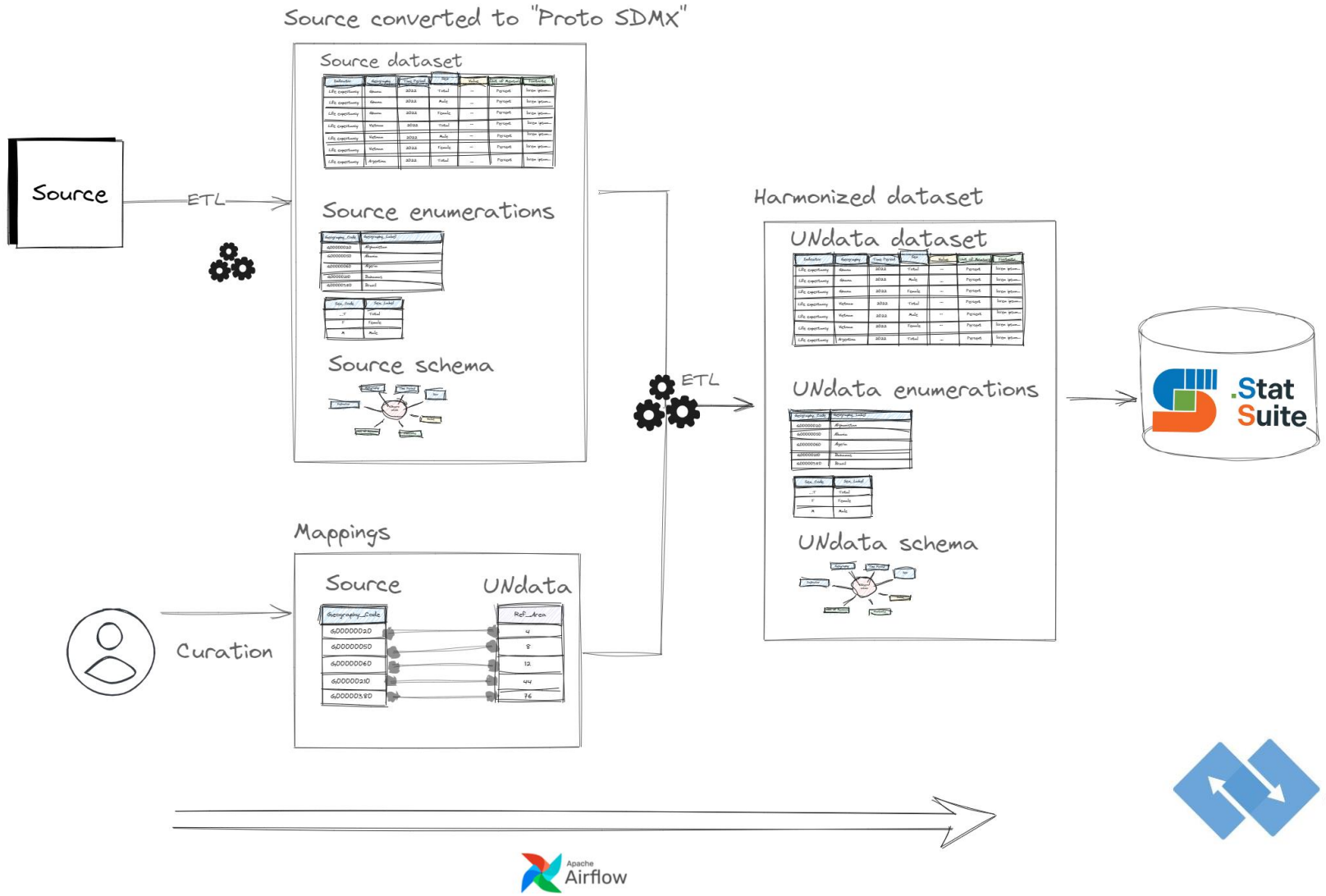
Ref_Area	Ref_Area_Desc
4	Afghanistan
8	Albania
12	Algeria
44	Bahamas
76	Brazil

UNDATA

United Nations | Department of Economic and Social Affairs

Data pipelines

Stage 1:

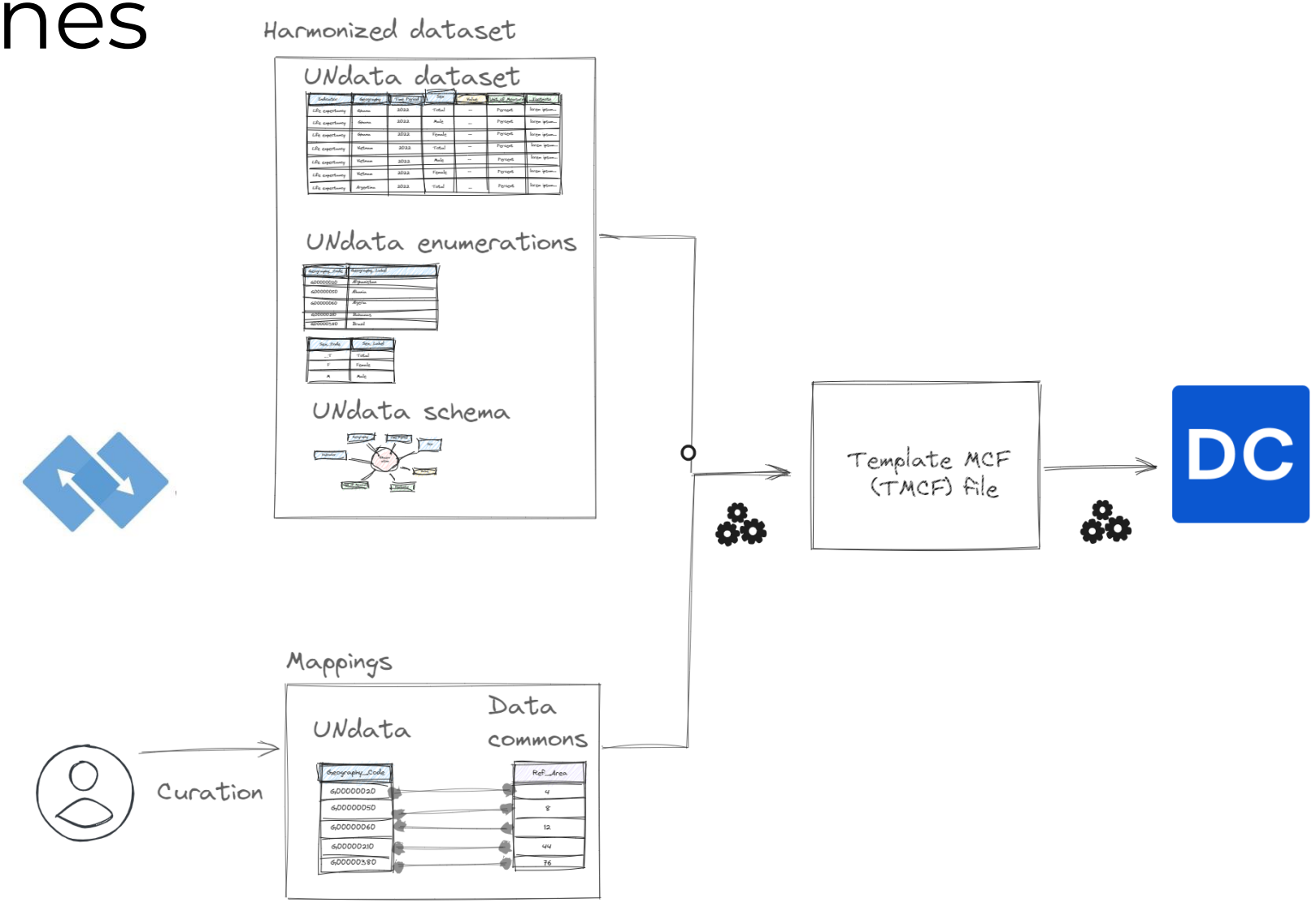


UNDATA

United Nations | Department of Economic and Social Affairs

Data pipelines

Stage 2:



UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons Schema

- The underlying data model for Data Commons is based on Schema.org.
- It models the world as a collection of entities, each having attributes and relationships with other entities.
- These entities fall under a taxonomy of types, and each entity is an instance of one or more types in this taxonomy.
- The conceptual foundation of the data model lies in knowledge representation systems commonly known as a "knowledge graphs"
- APIs like Node and SPARQL, as well as the Data Commons Graph Browser, provide access to the graph view.

DC

UNDATA

United Nations | Department of Economic and Social Affairs

Key features of Data Commons:

- Standard schemas and APIs allow to seamlessly integrate data from different sources, saving users from having to merge, clean, and standardize data by themselves
- Versatile Tool Suite available to any site publishing data using these schemas and APIs
- Collaborative ecosystem of interlinked instances that are built on open principles.
- A tool or application designed for one Data Commons is compatible with all, benefiting areas like visualization, analytics, machine learning, and natural language interfaces.

DC

UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons Schema

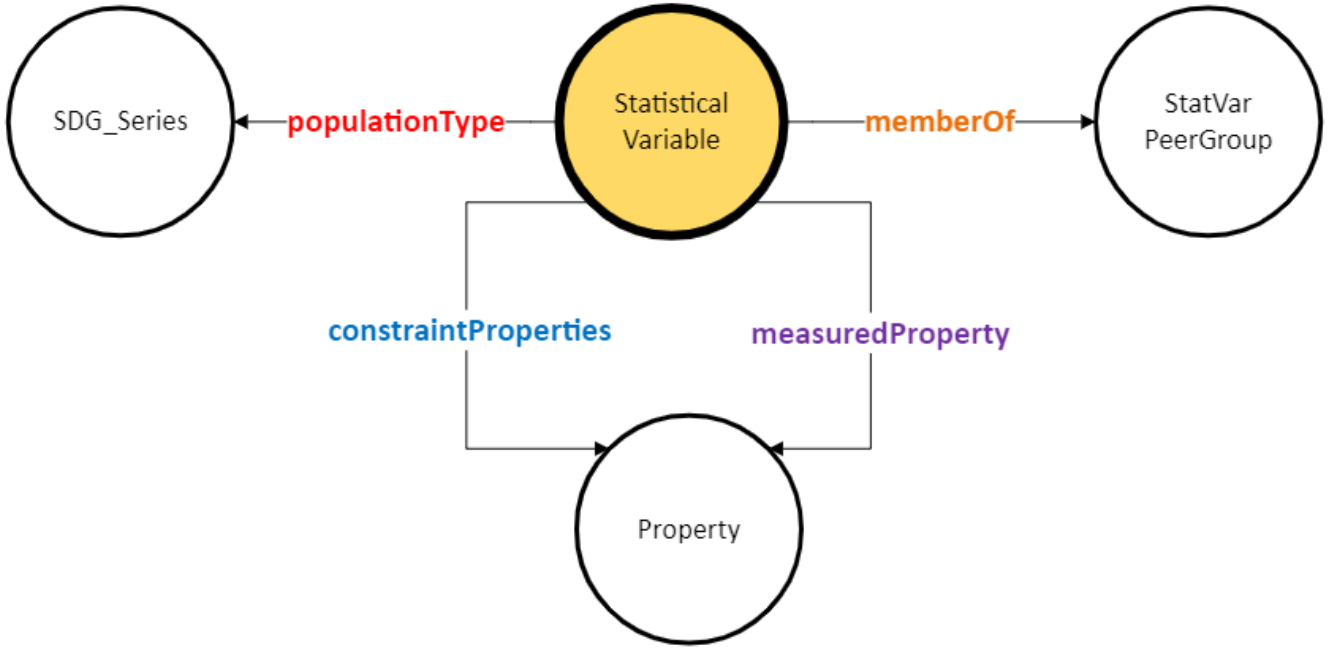
- The model captures the following elements of a data point:
  - **Entity**: The object or thing being measured → Place
  - **Variable**: The specific measurement being taken
  - **Observation**: The value of the variable for a given entity
  - **Provenance**: The source of the data

DC

UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons model for statistical variables

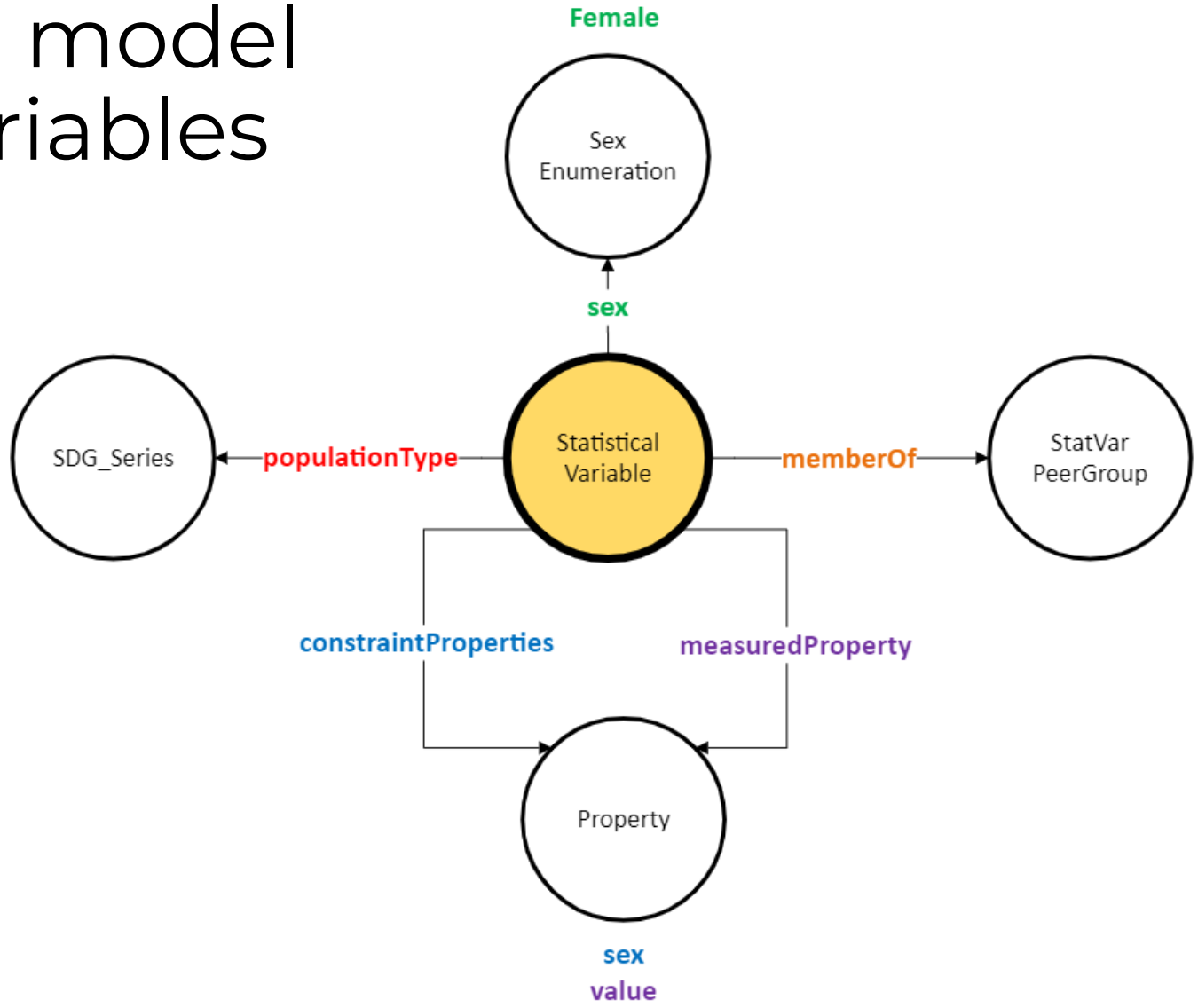


DC

UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons model for statistical variables

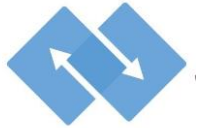


DC

UNDATA

United Nations | Department of Economic and Social Affairs

Contrast between SDMX Information Model and Data Commons schema



- Specifically designed for statistical data, with a more tabular, dimensional focus.
- More rigid but specific, focusing on dimensions, attributes, and measures.
- Uses StructureMaps and ComponentMaps to describe how data should be transformed or related.
- Better suited for statistical data where dimensions and measures are predefined.

- Based on the Schema.org model, with roots in knowledge representation systems.
- Aims for a more flexible, verbose base layer, allowing various kinds of relationships and attributes
- Provides different APIs (Node, SPARQL, DCGET) for various views, including a time-series view.
- Built as a knowledge graph, making it more suitable for capturing a wide range of relationships among diverse entities

UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons model for statistical variables

SDMX Series	SDMX slice definition	population type	statistical variable	constraint property	sex	member of
SI_POV_DAY1	SEX="M"	SDG_SI_POV_DAY1	**sdg/SI_POV_DAY1.SEX--M**	Sex	Male	dc/g/SDGSIPOVDAY1_sdgsex
SI_POV_DAY1	SEX="F"	SDG_SI_POV_DAY1	**sdg/SI_POV_DAY1.SEX--F**	Sex	Female	dc/g/SDGSIPOVDAY1_sdgsex
SI_POV_DAY1	SEX="_T"	SDG_SI_POV_DAY1	**sdg/SI_POV_DAY1**			dc/g/SDGSIPOVDAY1
	Type:	SDG_Series	**StatisticalVariable**	Property		StatVarGroup

DC

UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons for the SDGs

- 2023.Q3.G.01 data refresh (currently on the UNSD site)
- 2,172,161 StatVarObservations
- 3,807 StatisticalVariables
- 268,453 time series
- 6,467 geographies

UNDATA

United Nations | Department of Economic and Social Affairs

Data Commons for the SDGs

- UN SDMX → Data Commons MCF
- Each **observation** corresponds to one **StatVarObservation**
  - Associated with a specific StatisticalVariable, Place, and Date
- Each **SDG series** corresponds to 1+ **StatisticalVariables**
- Each **dimension** corresponds to one **Property**
  - Attached to the StatVarObservation or the StatisticalVariable
    (Properties help further specify StatVarObservations and StatisticalVariables)

UNDATA

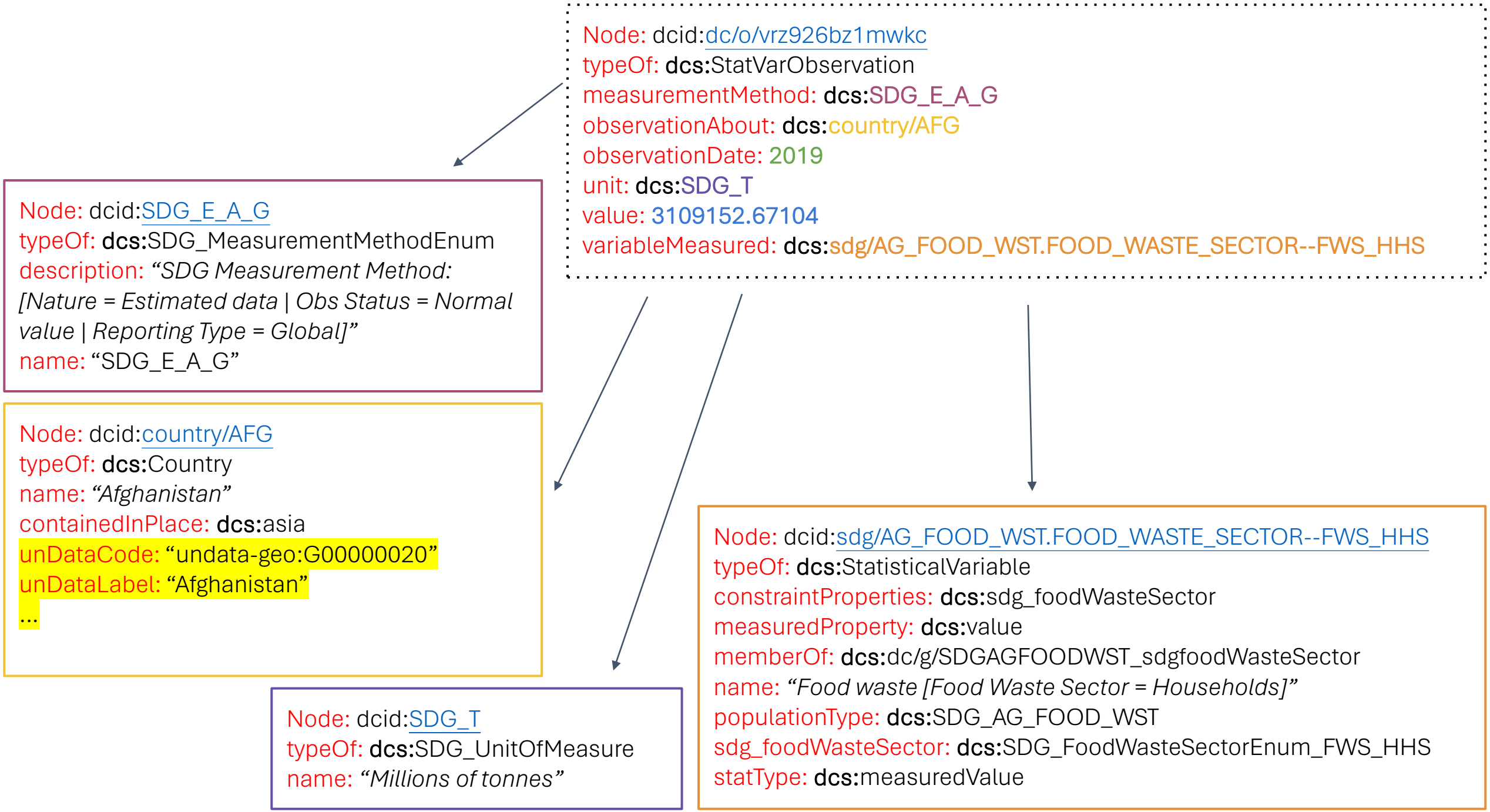
United Nations | Department of Economic and Social Affairs

UN SDMX → Data Commons MCF

SERIES_CODE,SERIES_DESCRIPTION,VARIABLE_CODE,VARIABLE_DESCRIPTION,VARIABLE_ACTIVE_DIMS,GEOGRAPHY_CODE,GEOGRAPHY_NAME,GEOGRAPHY_TYPE,GEO_AREA_CODE,GEO_AREA_NAME,CITIES,SAMPLING_STATIONS,IS_LATEST_PERIOD,TIME_PERIOD,TIME_DETAIL,TIME_COVERAGE,FREQ,FOOD_WASTE_SECTOR,OBS_VALUE,VALUE_TYPE,UPPER_BOUND,LOWER_BOUND,UNIT_MEASURE,UNIT_MULT,BASE_PERIOD,NATURE,SOURCE,GEO_INFO_URL,FOOT_NOTE,REPORTING_TYPE,OBS_STATUS,RELEASE_STATUS,RELEASE_NAME

AG_FOOD_WST, Food waste (Tonnes),AG_FOOD_WST@FOOD_WASTE_SECTOR--FWS_HHS,Food waste (Tonnes) [Food Waste Sector = Households],['FOOD_WASTE_SECTOR'],4,Afghanistan,Country,4,Afghanistan,,,True,2019,2019,,,FWS_HHS,3109152.67104,Float,,,T,,,E,Food Waste Index Report 2021 / WESR,,Very Low Confidence,G,A,Published,2023.Q2.G.01

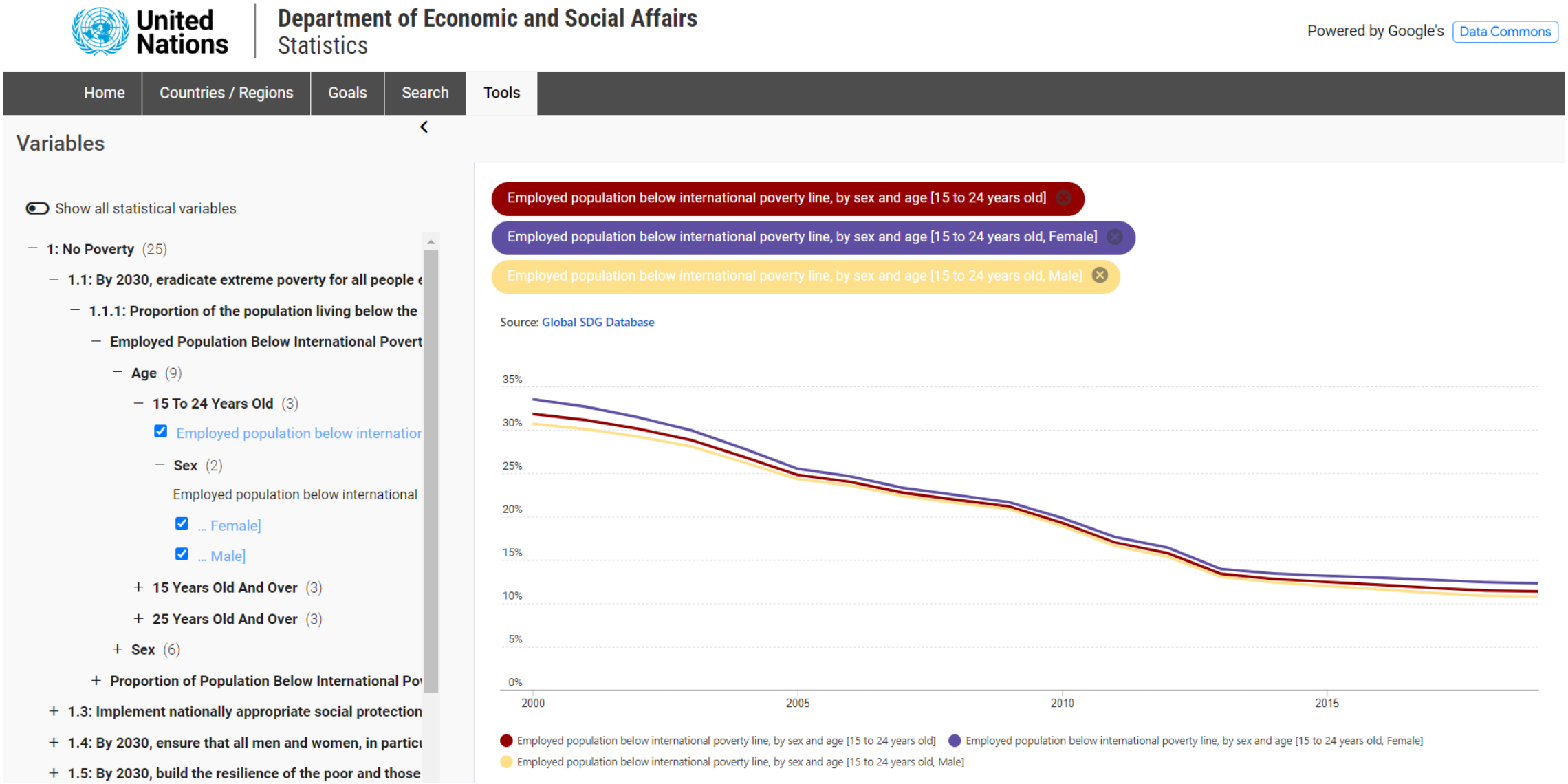
Node: dcid:dc/o/vrz926bz1mwkc
typeOf: dcs:StatVarObservation
measurementMethod: dcs:SDG_E_A_G
observationAbout: dcs:country/AFG
observationDate: 2019
unit: dcs:SDG_T
value: 3109152.67104
variableMeasured: dcs:sdg/AG_FOOD_WST.FOOD_WASTE_SECTOR--FWS_HHS


Node: dcid:dc/o/vrz926bz1mwkc
typeOf: dcs:StatVarObservation
measurementMethod: dcs:SDG_E_A_G
observationAbout: dcs:country/AFG
observationDate: 2019
unit: dcs:SDG_T
value: 3109152.67104
variableMeasured: dcs:sdg/AG_FOOD_WST.FOOD_WASTE_SECTOR--FWS_HHS
Node: dcid:SDG_E_A_G
typeOf: dcs:SDG_MeasurementMethodEnum
description: "SDG Measurement Method: [Nature = Estimated data | Obs Status = Normal value | Reporting Type = Global]"
name: "SDG_E_A_G"
Node: dcid:country/AFG
typeOf: dcs:Country
name: "Afghanistan"
containedInPlace: dcs:asia
unDataCode: "undata-geo:G00000020"
unDataLabel: "Afghanistan"
...
Node: dcid:SDG_T
typeOf: dcs:SDG_UnitOfMeasure
name: "Millions of tonnes"
Node: dcid:sdg/AG_FOOD_WST.FOOD_WASTE_SECTOR--FWS_HHS
typeOf: dcs:StatisticalVariable
constraintProperties: dcs:sdg_foodWasteSector
measuredProperty: dcs:value
memberOf: dcs:dc/g/SDGAGFOODWST_sdgfoodWasteSector
name: "Food waste [Food Waste Sector = Households]"
populationType: dcs:SDG_AG_FOOD_WST
sdg_foodWasteSector: dcs:SDG_FoodWasteSectorEnum_FWS_HHS
statType: dcs:measuredValue

Metadata Content Framework (MCF)

Node: dcid:sdg/SI_POV_EMP1. AGE--Y15T24
typeOf: dcs:StatisticalVariable
measuredProperty: dcs:value
name: "Employed population below international poverty line, by sex and age [15 to 24 years old]
populationType: dcs:SDG_SI_POV_DAY1
statType: dcs:measuredValue
age: dcs:SDG_AgeEnum_Y15T64

https://browser.datacommons.org/browser/sdg/SI_POV_EMP1.AGE--Y15T24



Goals, **Targets**, and **Indicators** are specified UNSD. The remaining sub-hierarchy is auto-generated from the data.

UNDATA

United Nations | Department of Economic and Social Affairs

How to specify a new geography?

- Each place needs a unique DCID, but can have arbitrary additional Properties, such as the UN-specific unDataCode and unDataLabel
  - “United States of America”
    - dcid: country/USA
    - unDataCode: undata-geo:G00003340
    - unDataLabel: “United States of America”
    - …
- Each place specifies its containment via the **containedInPlace** Property
  - “United States of America” (dcid: country/USA) **containedInPlace** “North America” (dcid:northamerica)
  - “North America” (dcid:northamerica) **containedInPlace** “Earth” (dcid:Earth)

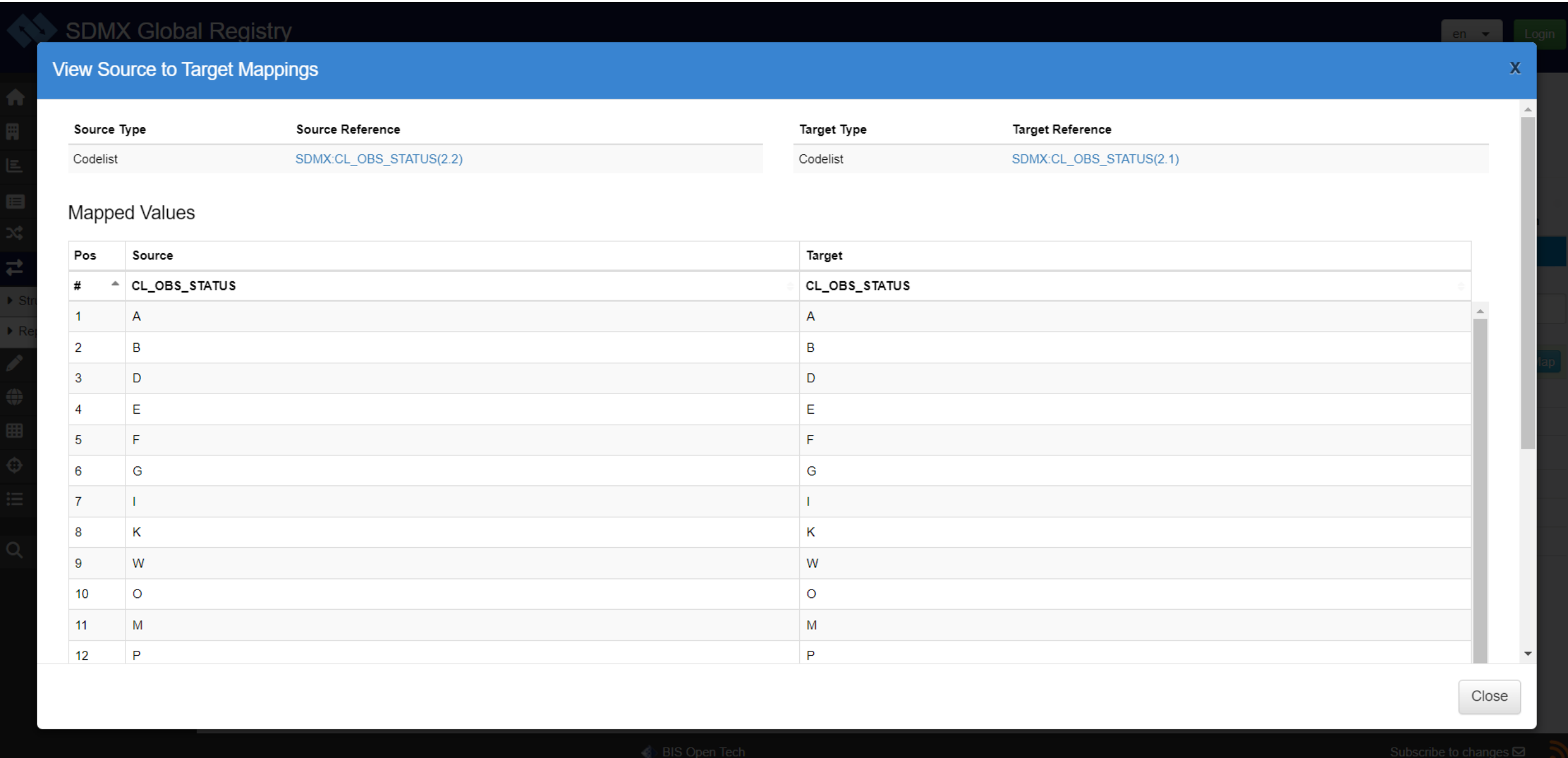
UNDATA

United Nations | Department of Economic and Social Affairs

Challenge: Make the ongoing maintenance of mappings scalable

- To scale to real-world use cases, automated tools are critical, e.g.:
  - Ontology matching
  - Entity resolution
  - Semantic similarity
  - Automated reasoning
- No single tool will perform equally well on all inputs
- Purely automated mappings often need to be refined by hand or using sophisticated reconciliation approaches
  - Develop tools to help **understand schemas and maintain data mappings** for each source
  - Employ machine learning or AI to recognize patterns and assist in data mappings???

SDMX Representation Maps



SSSOM

- Rich and **extensible vocabulary** for describing many types of individual mappings and sets of mappings
- Includes simple **tabular format** for dissemination of mappings
- **Community-driven** standard with well-defined governance and collaborative workflows
- Maintained as an **open-source** project on the mapping-commons GitHub organization (https://mapping-commons.github.io/sssom/spec/)
- Specification available from: https://w3id.org/sssom/
- Many SSSOM are mapped to external vocabularies (e.g., PAV, Dublin Core, …)

UNDATA

United Nations | Department of Economic and Social Affairs

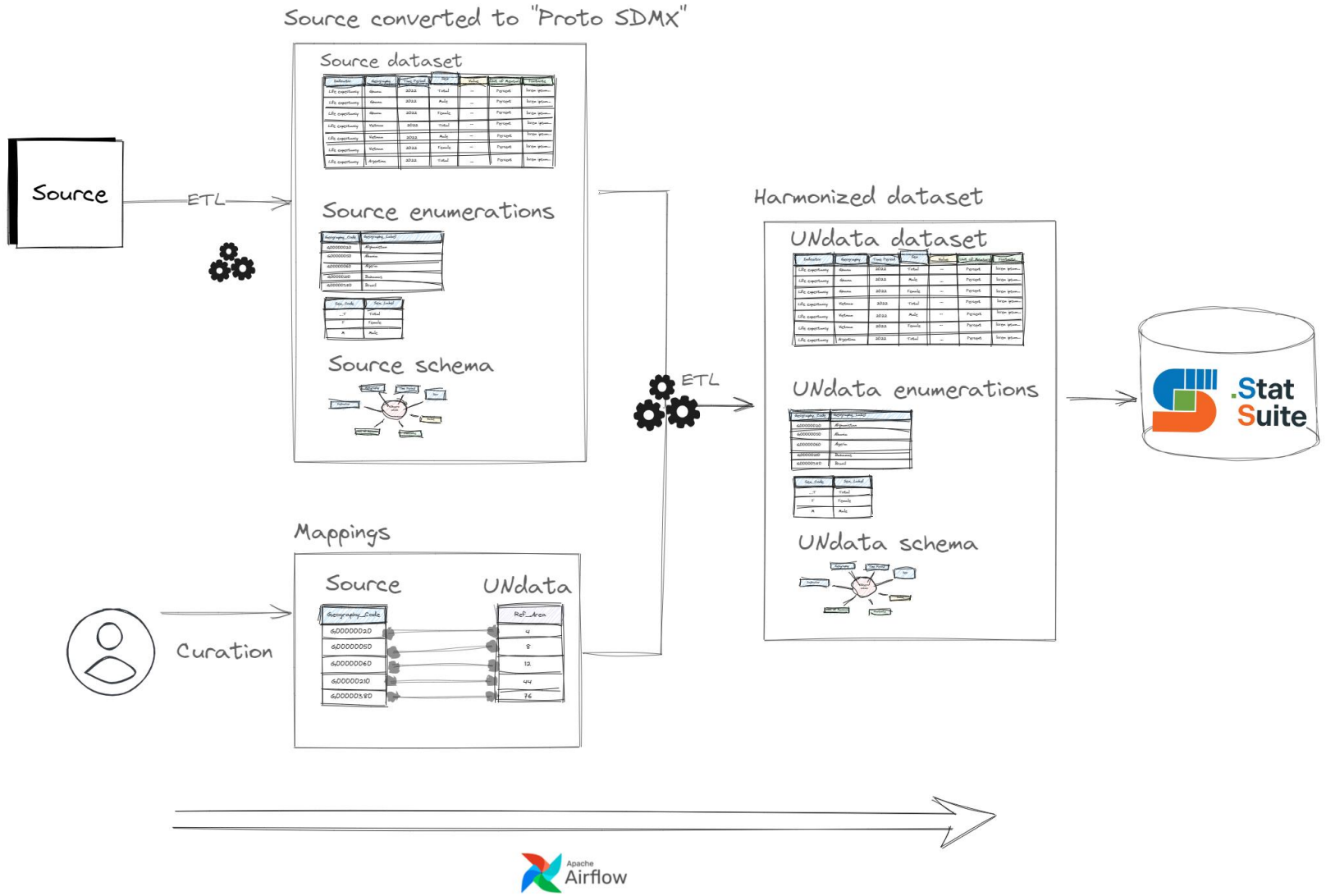
Challenge: Make data pipelines observable and adaptable

- Implement dashboards to **visualize ETL flow** and track data movement
- Integrate **automated validation checks** throughout the ETL process
- Set up **alerts for any discrepancies** or issues in the data processing
- Create flexible **connectors** for various data sources to ease integration.
- Maintain comprehensive **documentation** on data transformation rules and logic

UNDATA

United Nations | Department of Economic and Social Affairs

Data pipelines



United Nations | Department of Economic and Social Affairs