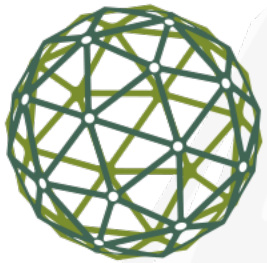


8th International Conference on BIG DATA & Data Science for Official Statistics

BILBAO 2024

Informing Climate Change and Sustainable Development Policies with Integrated Data

BILBAO. SPAIN | 10-14 JUNE 2024 | #UNBigData2024

UN Data and the Use of Data Commons Infrastructure

United Nations BigData

EUSKO JAURLARITZA GOBIERNO VASCO

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS EXTERIORES, UNIÓN EUROPEA Y COOPERACIÓN

BC3 BASQUE CENTRE FOR CLIMATE CHANGE Klima Aldaketa Ikergai

EXCELENCIA MARÍA DE MAEZTU 2023-2027

INE Instituto Nacional de Estadística

Eustat EUSKAL ESTATISTIKA ERAKUNDEA INSTITUTO VASCO DE ESTADÍSTICA

UN Data

Powered by Google's Data Commons

UN Data

Powered by Google's Data Commons

The UN Data modernization project

- 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

Powered by Google's Data Commons

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

UN Data

Powered by Google's Data Commons

Concrete application of CDIF* 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 web infrastructure

*Cross-Domain Interoperability Framework https://doi.org/10.5281/zenodo.11236871

UN Data

Powered by Google's Data Commons

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 as dimensional data repository, leveraging SDMX standards
- Deploy on UN Global Platform infrastructure
- Engage with statistical community to collaboratively manage data integration process

UNData
Powered by Google's Data Commons

Towards a distributed UNdata Knowledge Graph

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

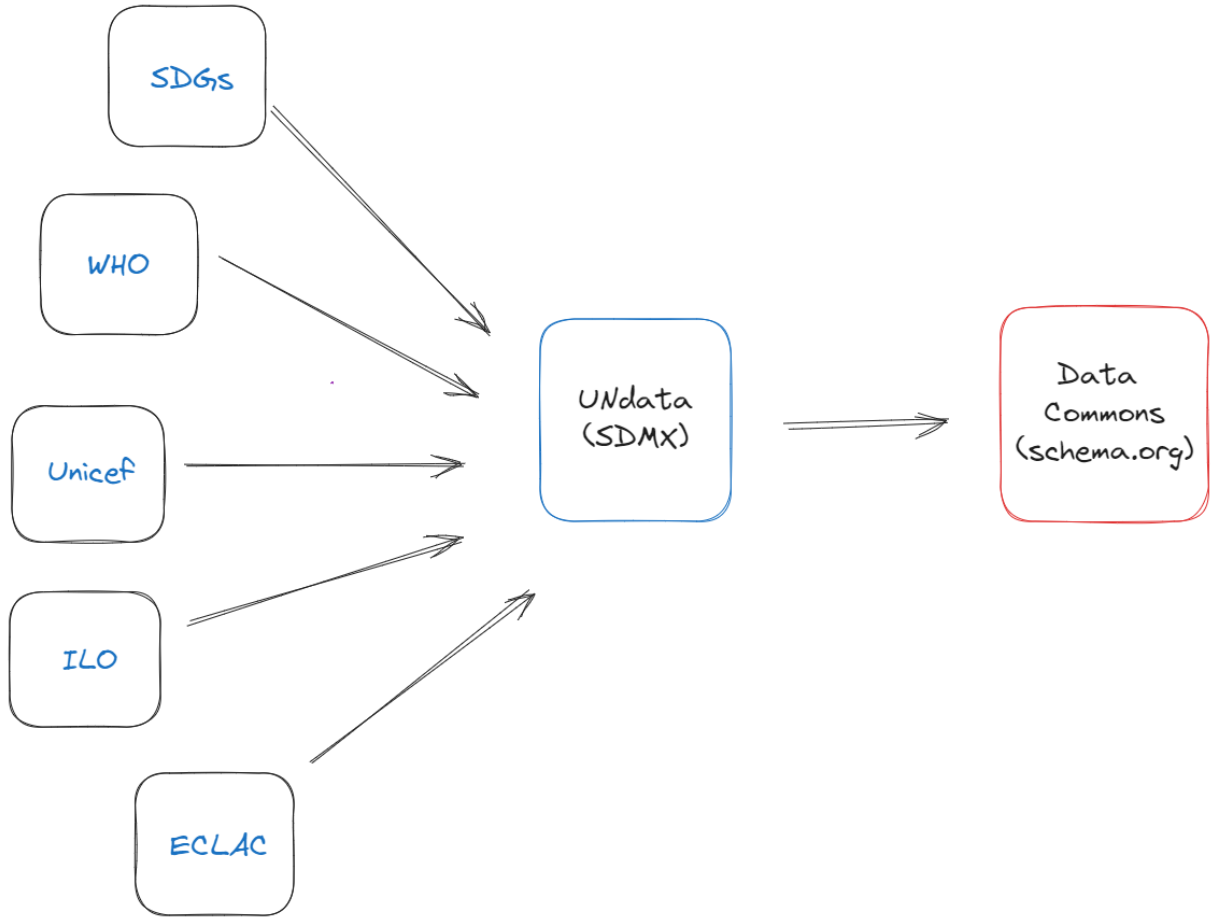
- Establish explicit and implicit links to external resources, thus making data more easily findable, searchable, and usable.
- Build applications that efficiently access related data across multiple domains using linked open data techniques
- Generate insights by reasoning over complex relationships.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

UN Data

Powered by Google's Data Commons

Objective

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



UNData
Powered by Google's Data Commons

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.

UN Data
Powered by Google's Data Commons

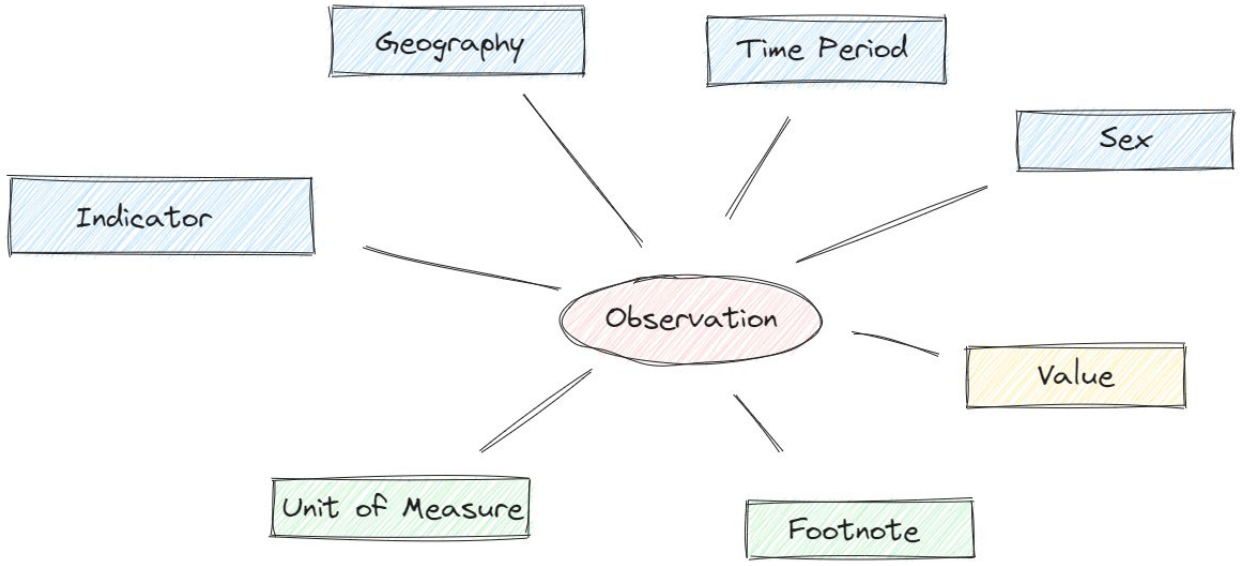
Data integration problem

- Equivalent concepts (entities) are assigned different identifiers in different databases or vocabularies
- Different agencies use different names for the same real-world entities
- A significant part of the work consists in aligning terms that refer to the same concept across datasets.
- 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

UN Data
Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs

Concept schemas



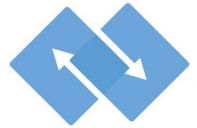
UN Data
Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs

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



UN Data
Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs

Mappings between source entities and UN Data entities

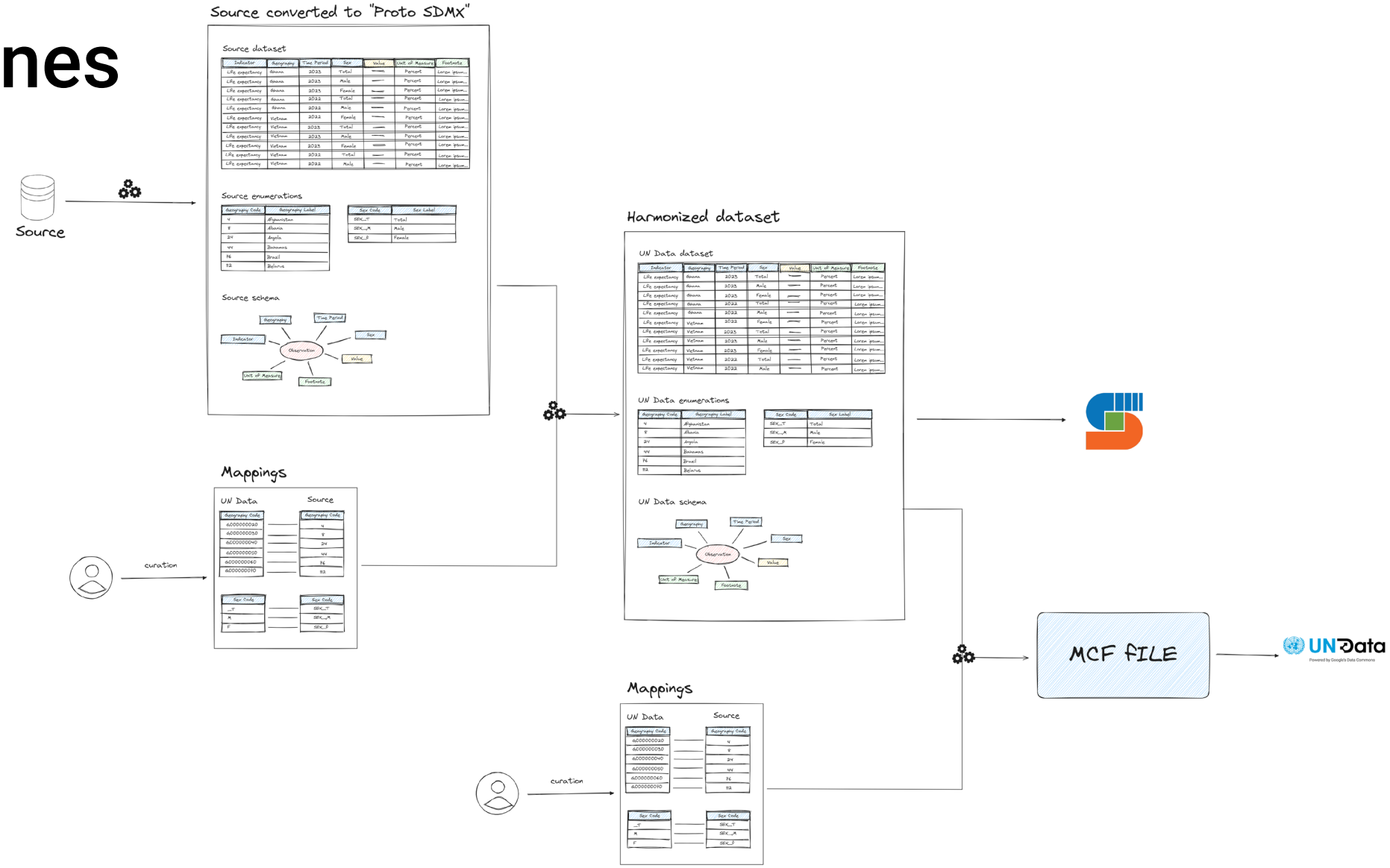
Geography_Code	Geography_Label
G00000020	Afghanistan
G00000050	Albania
G00000060	Algeria
G00000210	Bahamas
G00000380	Brazil

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

UNData
Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs

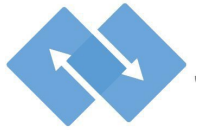
Data pipelines



UN Data
Powered by Google's Data Commons

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

UN Data
Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs

Additional data structures

- Taxonomy of UN Data themes and sub-themes

- Definition of statistical variable groupings

UN Data
Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs

Enriched metadata

Additional metadata is captured for each indicator or sv-grouping:

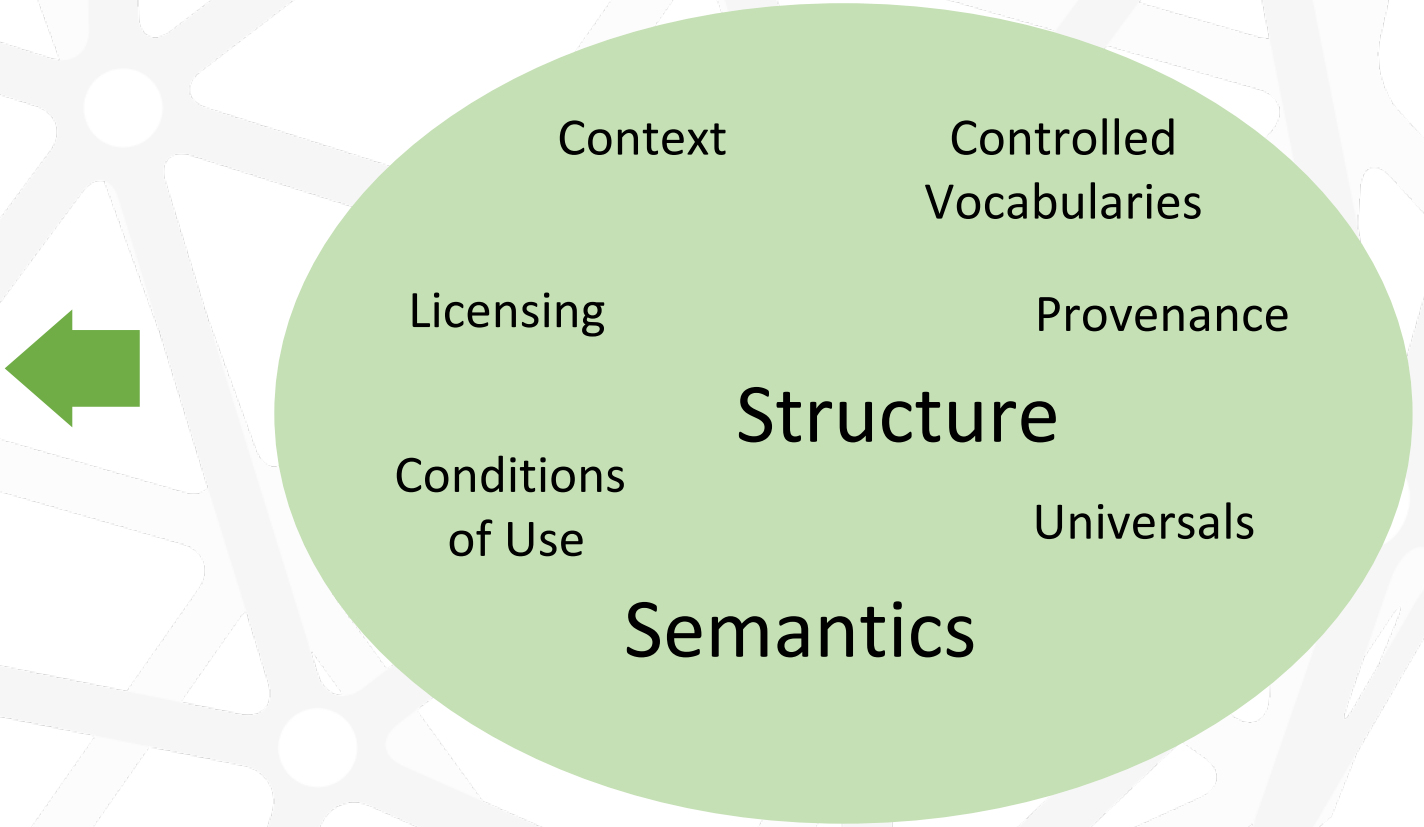
- Metadata link
- Revised sv-group description

DC

UN Data
Powered by Google's Data Commons

Functions of Cross-Domain FAIR implementation

- Discovery
- Assessment
- Access
- Integration
- Packaging



UNData
Powered by Google's Data Commons

Alignment with CDIF recommendations

- Re-express domain metadata (e.g., SDMX) into other common, integration-ready data descriptions (e.g., schema.org).
- Manage data at the finest granularity possible to facilitate composability
- Use persistent, de-referenceable identifiers at all levels of granularity

UN Data

Powered by Google's Data Commons

Making UN Data AI-ready

- Data profiling for **Discovery**, **Assessment** and **Access**
- Mappings of concepts and code lists for **Integration**
- Supplementary metadata for **Context** and **Provenance** traceability (e.g., edited indicator descriptions, mapping to UN thematic areas, definition of variable groupings, methodological notes…)

United Nations | Department of Economic and Social Affairs
Statistics

UNData
Powered by Google's Data Commons

UNData

Powered by Google's Data Commons

How much food goes wasted around the world?

Introducing UN Data, a powerful tool for extracting insights from data available across the UN system. Search and explore high-quality datasets and digital public goods right at your fingertips to empower and expedite evidence-based decision-making.

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

UNData

Powered by Google's Data Commons

United Nations | Department of Economic and Social Affairs