United Nations | Department of Economic and Social Affairs

AI -readiness and interoperability: the case for unique identifiers

Luis G. Gonzalez Morales, Chief, Data Innovation Section, UNSD

Side event:

World Geographic Names Database and a system of unique identifiers for cities

UNGEGN Working Group on Geographical Names Data Management

1 May 2025

Background: Linked Open Data in UNSD

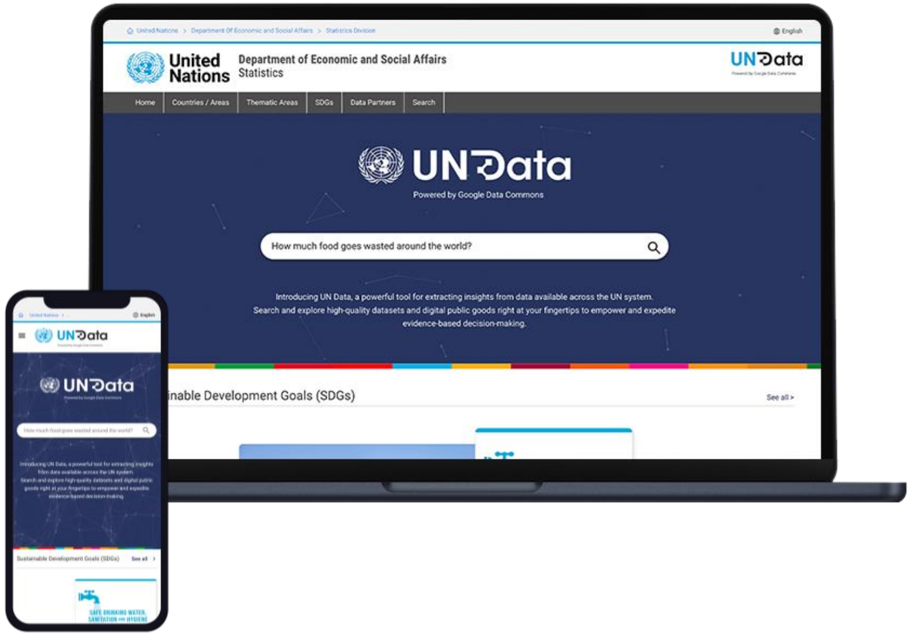
- To support the integration of data and information to monitor and implement the 2030 Agenda for Sustainable Development, UNSD developed a system of common Internationalized Resource Identifiers (IRIs) for Goals, Targets, Indicators and related statistical series.
- The Secretariat of the High-Level Committee on Management (HCLM) of the UN System Chief Executives Board for Coordination has encouraged UN system organizations and relevant stakeholders to map their SDG-related resources to the SDG Internationalized Resource Identifiers.
- Since then, the Statistics Division has continued to work, in collaboration with partners from across the UN System, in the development of Linked Open Data infrastructure for the integration and dissemination of official data and statistics

UN Data
Powered by Google Data Commons

United Nations | Department of Economic and Social Affairs

UN Data Modernization

- Make UN Data AI-ready by harmonizing concepts for semantic integration
- Organize data in extensible, machine-actionable structures.
- Support natural language search using LLM technologies
- Enable scalable integration of datasets across domains and organizations.


United Nations > Department Of Economic and Social Affairs > Statistics Division
English
United Nations
Department of Economic and Social Affairs
Statistics
UN Data
Home
Countries / Areas
Thematic Areas
SDGs
Data Partners
Search
UN Data
Powered by Google Data Commons
…inable Development Goals (SDGs)
See all >
Sustainable Development Goals (SDGs)
See all >
SAFE DRINKING WATER, SANITATION AND HYGIENE

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.

Caliper and LOD in the UN Global platform

- An infrastructure and a set of methodologies for the FAIR management and dissemination of statistical classifications along their entire lifecycle with a Link Open Data approach
- An international partnership to manage statistical classifications as Public Goods to support official statistics
- Originally developed at FAO, Caliper infrastructure is now hosted in the UN Global Platform in the Aries for SEEA Thematic Hub, for a 2-year pilot with the UNSD.

“Identifiers are part of the basic data infrastructure of any data ecosystem. They help link data across different domains or communities, enabling cross-disciplinary work and collaboration.”

The unique identifier problem

- Equivalent concepts (entities) are assigned different identifiers in different databases or vocabularies
- Different communities use different names for the same real-world entities
- Integrating datasets that follow different identifier systems is costly (regardless of whether this is done manually by experts or by automated tools)
→ But the cost of incorrect or incomplete data linking can be even higher!
- Linking 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

Linking different geographic names databases

- Similar lists of geographic names are often developed by different organizations, for different purposes, or evolve across versions over time.
- Correspondences can exist between different gazetteers, naming conventions, or individual place entries:
  - Between gazetteers covering the same domain (e.g., national versus international place name registries)
  - Between different but linked domains (e.g., administrative boundaries and populated places)
  - Historical correspondences (e.g., former place names to current official names)
  - Versioning of entries over time within a given gazetteer or naming system

Definition of mappings

- Mapping is synonymous to “**correspondence**” in the ontology matching literature
- A mapping is a `<subject, predicate, object>` triple that reflects a correspondence between the representations of two real-world entities living in two different data spaces
    - The **subject** and **object** are the two entities being mapped
    - The **predicate** defines the type of relationship between them, such as:
        - “Exact match” (skos:exactMatch)
        - “Equivalent class” (owl:equivalentClass)
        - …
- A special kind of mapping set is an “**alignment**” comprising all mappings between two data spaces (ontologies/databases)

Lack of standardized representation

- Most mappings are often represented as two-column tables with matching terms, or as cross references in one of the two datasets being mapped
- Semantics about the nature of the mappings are often missing:
    - Exact
    - Broader/Narrower
    - Closely related but neither exact nor broader/narrower
    - Etc.
- Provenance is often missing (was the mapping reviewed by a domain expert?)

→ This makes it very difficult to reuse mappings and combine mappings from different resources

Linking decisions and degree of confidence

- Linking decisions need to be **explainable**
- Human curators often have different levels of confidence about the accuracy of any given mapping
- Different use cases may require different levels of mapping precision
  - Entity merging and data translation / transcoding require exact matches
  - Data grouping may only require broad/narrow matches
  - Machine learning use cases can benefit from close and related matches, regardless of their lack of semantic precision

Linking tools

- Challenge: Make the ongoing maintenance of mappings scalable
- To scale to real-world use cases, automated tools are critical
- Automated matching techniques include:
  - Entity resolution (the task of determining whether two database records correspond to the same entity),
  - Semantic similarity
  - Automated reasoning
- Recent semantic-aware AI/ML approaches allow to
  - work with messier inputs
  - exploit the metadata structures to determine high-quality
- Purely automated mappings often need to be refined by hand or using sophisticated reconciliation approaches

Identifiers best practices

- **Unique** (i.e., no two objects should have the same identifier, and an object should have only one identifier).
- **Universal** (i.e., every object must have an identifier).
- **Immutable** (i.e., an object should have the same identifier over its entire lifespan).
- **Never re-issued** (i.e., once an identifier has been assigned to an object, it should not be re-assigned to another object, even if the original object ceases to exist).
- **Well-documented** (so it is properly understood and used across systems)
- **Issued from a central authority** (to ensure it is consistently assigned and managed across systems).
- **Meaning-independent:** (i.e., the characteristics of the identified objects should be presented to the user as a separate attribute with its corresponding user-friendly textual description).
- **Available to and accessible by all relevant systems** (i.e., identifiers should, as a general rule, be available to third-party applications)

Linked data principles

- Best practices for publishing and connecting structured data on the web so that it can be easily discovered, linked, and reused across different systems
  - Use URIs to identify things
  - Use HTTP URIs so that people (and machines) can look them up
  - Provide useful information when a URI is looked up, using open standards
  - Include links to other URIs to enable discovery of related data
- In geographic names or gazetteers, applying linked data principles would mean:
  - Each place has a stable URI
  - Metadata about the place is structured and retrievable
  - The place is linked to other datasets (like historical records, statistical data, administrative boundaries)

Uniform Resource Identifieres

- URIs are standardized strings of characters that identifies a resource uniquely and unambiguously on the Internet.

  [scheme]://[host]/[path]/[local identifier]

- Compact URIs (CURIES) consist of a registered prefix (namespace) and a locally unique identifier (reference), separated by colon (:).

  The prefix (namespace) is mapped to the [scheme]://[host]/[path]/ part of the URI / IRI.

- Example:

  URI = https://undata.org/geo-1

  CURIE = undata:geo-1

Unique identifiers in the era of AI

A system of Unique Identifiers can help operationalize the recommendations of the report on the Role of AI in evidence-based geographical names management

- Enable consistent, evidence-based and AI-assisted processing of toponyms across languages, formats, and contexts
- Help AI systems disambiguate between places with identical names
- Provide a persistent reference point across datasets, enabling AI-systems to cross-reference historical, linguistic, and administrative data with high precision
- Improve quality and consistency of AI-generated results
- Foster interoperability between national and international gazetteers, statistical databases, and AI training corpora

How could the Secretariat support the initiative?

- Facilitate consultative technical process with Member States and relevant bodies
- Coordinate alignment with existing standards
- Serve as central documentation and dissemination hub
  - Disseminate unique identifiers in Caliper
  - Implement unique identifiers in World Geographic Names Database
- Promote adoption and capacity building
  - Develop and maintain technical documentation
  - Coordinate the creation of training materials, templates, and guidance notes