Bringing statistical classifications to the semantic web

Needs of the statistical community

- Organizing the collection, processing, analysis and dissemination of statistical data
- Harmonizing statistical concepts
  → define once; re-use multiple times.
- Making statistical data FAIR

Data integration challenge

- The data landscape is:
  - **Highly decentralized**: With different actors responsible for different domain-specific datasets
  - **Diverse**: With wide variety of domains, each using specialized terminology, formats, levels of disaggregation…

Statistical Classifications

- Enable the collection, organization, analysis, and dissemination of statistical data
- Facilitate clean comparability and computability of data
- Enable people and machines to find, access, and integrate diverse datasets

The nature of the web

- The web is a distributed, open community – anyone can publish any data and content for everyone to see.
- There is no one authority single-handedly ruling over all additions or modifications to the data that is available online
- When the information from two sources on the web is brought together, there is no a priori way of knowing whether they refer to exactly the same thing, to two completely different things, or somewhere in between.

Linked Open Data

- Technique for organizing data on the web
- Powerful mechanism for data integration
- Allows to find and combine related data and metadata across domains in a meaningful manner
- Enabler of:
  - Open data
  - FAIR principles

Bringing statistical data to the semantic web

- Linked Open Data principles are increasingly embraced by the statistical community
- Growing collection of tools:
  - Data Cube Vocabulary
  - StatDCAT Application Profile
  - Core Ontology for Official Statistics
  - XKOS - An SKOS extension for representing statistical classifications
  - Caliper - A tool developed by FAO for the dissemination of statistical classifications in LOD format

Interoperable Classifications for Semantic Web

- Making statistical classifications interoperable enhances their utility and reach.
- When classifications are available in LOD format, it is possible to automate data integration and analysis tasks, and to make statistical data easier to find and to access.

Uniform Resource Identifiers (URIs)

- In the Semantic Web, every resource has a URI
- Powerful, standard way of creating **globally unique** identifiers
- Similar to how webpages are uniquely identified using URL addresses
- URIs used to identify data resources on the web must comprise of:

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

- Example:

  "https://unstats.un.org/classifications/ISIC/rev4/"

  "https://unstats.un.org/classifications/ISIC/rev4/C"

Role of URIs in making statistical data FAIR

- **Findable**: URIs act as specific "addresses," making data easier to locate.
- **Accessible**: Once found, the URI enables a user to access the data directly.
- **Interoperable**: URIs provide a standard format that allows data to be easily integrated with other datasets, supporting machine-to-machine communication.
- **Reusable**: The URI can be shared and reused, both by humans and computers, for various analytical and informational tasks.

Resource Description Framework (RDF)

- Provides the foundation for the implementation of Linked Open Data on the Semantic Web
- Based on the spirit of re-use and extension
- Simple “graph” model consisting of pairs of interconnected “triples” (subject-predicate-object).
- RDF and URIs together allow for precise identification and description of virtually anything, making data more organized and useful.
- RDF allows to meaningfully merge different data models and to view the information represented in them as a single, unified whole

SKOS

- Designed to represent “concepts” alongside their definitions, labels and notations (codes), as well as **basic** semantic relationships between concepts
- Widely used to represent taxonomies, thesauri, and other knowledge organization systems in machine-readable way

XKOS

- Extension of SKOS designed to meet the needs of the official statistics community
- Designed to provide more nuanced types of hierarchical relations between concepts that are required by statistical classifications.
- Borrows from the Neuchâtel model, with some modifications
- Allows to represent statistical classifications with all their structure and textural properties, as well as relations between classifications (and classification versions)

Why Linked Open Data?

- By publishing data in LOD format allows AI and data science applications to leverage knowledge about temporal, geospatial and logical relations
- For example: Ensuring that entities which are mutually exclusive are treated as such (a product cannot be a country; a UN Member State cannot be a UN Observer State, …)

Example:
Towards a distributed UNdata Knowledge Graph

The objective is to use semantic web technologies 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.

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

Example:

ARIES for SEEA uses AI to automate data and model integration

- The use of semantic web technologies, including statistical classifications in LOD format, allows the system to:
  - Select domain, location and time-specific resources,
  - Automatically choose the most suitable model based on user-defined criteria.
- Users can add their own data and models to ARIES-compatible networks, enriching the resource pool for future use.
- The interlinked nature of LOD allows for efficient querying and filtering, ensuring most relevant datasets and models are selected.

Semantic modeling is not easy

- Software developers and data engineers tend to under-specify meaning when building data models
- Ontologists, linguists, and domain experts tend to over-specify meaning and to debate semantic distinctions that users don’t care about

References

- P. Alexopoulos (2020). Semantic Modeling for Data
- D. Allemang and J. Hendler (2011). Semantic Web for the Working Ontologist, 2nd Ed.
- F. Cotton, D.W. Gillman, Y.Jaques (2013). Extended Knowledge Organization System (XKOS).