Improving semantic interoperability and discoverability of Official Statistics: A pseudo Knowledge Graph to bridge the semantic gap

J.Grazzini, M. Meszaros, J.-M. Museux and A. Skibinski

Eurostat, European Commission

Keywords: Structured data & metadata; Text & unstructured data; Semantics; Knowledge Graph; Linked Open Data; Machine learning & Natural language processing.

1. INTRODUCTION: CONTEXT AND OBJECTIVES

Although statistical (open) data is heavily used by numerous actors (*e.g.* businesses, academia, data journalists and other organisations) and for several purposes (*e.g.* policymaking, research and development), its potential for value creation grows when it is combined with other data [1]. The reuse of statistical data is however hampered by semantic interoperability challenges [2], *i.e.* challenges related to the interpretation of the meaning of data and metadata coming from different sources, and the integration of data of different types and formats. In a typical scenario, a statistical office publishes the statistical data on its portal[1] along with metadata (describing the structure of the data), which allow users to identify, access and consult statistical data, but also understand the quality, the formats, the collection mode *etc*... Nevertheless, the implementation of advanced data services, including faceted search, guided query builders, as well as services for data exploration and visual data browsing, is required to further support users when accessing and consulting data[2].

In this paper, we specifically focus on facilitating search and discovery, as well as sharing and reuse of information, without addressing issues of interoperability, namely:

- semantic search that provides users with flexible and faster access to the data through the ability to use natural language to query "things", *e.g.* concepts, data, categories, with both unstructured and structured content.
- discovery of (somehow hidden) facts and relationships based on navigation through hidden patterns and inferences that allow for large-scale analysis and identification of related things by users.

In practice, we consider the most common semantic architecture to represent robust actionable knowledge in the form of a *Knowledge Graph* (KG): we are indeed concerned not only with the structure of data but also with its meaning. For this purpose, we complement the (structured) information provided by metadata with the (semi-structured) information available online from which textual information can be extracted.

We conducted a proof of concept as a prototype in a test environment based on selected datasets from Eurostat data portal and textual material from Eurostat online website. We built a pseudo-KG that holds just enough information to make some of these resources findable and searchable, but that also contains the metadata to identify a particular representation of these resources in the traditional database. An early version of the

[1] Say, for instance, Eurostat online database: https://ec.europa.eu/eurostat/data/database.
[2] Through, for instance, EU Open Data portal: https://data.europa.eu/euodp/en/data/publisher/estat.

prototype can be explored online[3]. In the long term, this approach should provide with the foundation for making data discovery and access more intuitive and semantically relevant. Ultimately, it should help users to discover things they would have never gotten exposed[4] to before and further engage them into an interactive dialogue with the data.

2. THE TEXTUAL INFORMATION IS A RICH (META)DATA ASSET

Given the increasing values of organisational knowledge in the marketplace nowadays, we can argue that it is necessary to treat all sources of information and data as assets and try to optimise related benefits. From that perspective, the data and information available in textual form at Eurostat – in large document archives, in news releases, in scattered webpages, in scientific and technical publications, *etc*... – including its online website[5], is an invaluable source of knowledge[6]... but it is yet hard to access.

As a source of controlled vocabularies, Eurostat *Thematic Glossaries*[7] are a useful material for making data content searchable and findable. The knowledge exposed on these pages has been collected from experts in the various statistical domains. It is mutually exclusive and collectively exhaustive by design[8]. It is organised as a hierarchical structure that represents concepts in different levels of granularity so that we can relate higher-level concepts, like *"living conditions"*, to very specific ones, like *"equivalised disposable income"* by simply mapping the weblinks.

To build the pseudo-KG, we make practical use of the taxonomy provided by *Thematic Glossaries* pages as well as the textual information availabble in *Statistics Explained*[9] articles. Since they include additional information about the usage of each term and about its relationship to broader, narrower, related and or equivalent terms – similar to the *Simple Knowledge Organization System*[10] (SKOS) ontology – they help ensure consistency and avoid ambiguity in the description of data.

3. A NAIVE GRAPH IS A FIRST INSTANCE OF THE KNOWLEDGE GRAPH

While early semantic challenges are met by existing approaches, *e.g. Linked Open Data* technologies[11], other technologies have emerged [4][5][6] that make possible to infer knowledge bases on massive text corpora without necessarily relying on a database that has already been annotated by a human. The process of knowledge extraction into the KG is in this case significantly enhanced by the ability to automatically establish meaningful links between textual sources of information.

By combining *Web Scraping* (WS) tools together with *Natural Language Processing* (NLP) techniques, a pseudo-KG is built, backed by a graph database and a linked data store, providing the platform (and interface) required for storing, reasoning, inferring, and using data with structure and context. The nodes in the graph refer to any type of

[3] **http://63.34.157.226/home**

[4] *E.g.*, being able to see the needed information without having to explicitly type it into a search bar.
[5] Actually, in terms of quantity, most of the information and data is textual.
[6] From a corporate perspective, *Eurostat* available text material has been already recognised as invaluable resources to enrich in-house knowledge [3].
[7] https://ec.europa.eu/eurostat/statistics-explained/index.php/Thematic_glossaries
[8] In principle. In practice, some duplication/redundancy is actually observed.
[9] https://ec.europa.eu/eurostat/statistics-explained
[10] https://www.w3.org/TR/skos-reference/
[11] https://ec.europa.eu/eurostat/web/nuts/linked-open-data

"things" – *e.g.* concepts, data, and categories – that are extracted and filtered from webpages using NLP. The edges of the graph represent the relationships that exist between any of the things in the graph, as long as these things appear in the webpages that are linked and retrieved using WS (see Figure A). The resulting graph shares a pseudo ontological schema similar to SKOS and provides with a systematic approach to storing expert information about both statistical datasets and statistical concepts, and the relationships between them (see Figure B). Semantic similarity searches are based on embedding of relationships (connections) in the graph. This low-level processing is enough to help people find relevant information (content of nodes) through easy traversal along the relationships (links through edges). It also provides with the scalability and flexibility needed to expand categorisation to any number of things (by increasing the number of nodes) and represent the relationships from any number of different data (by increasing the number of edges).

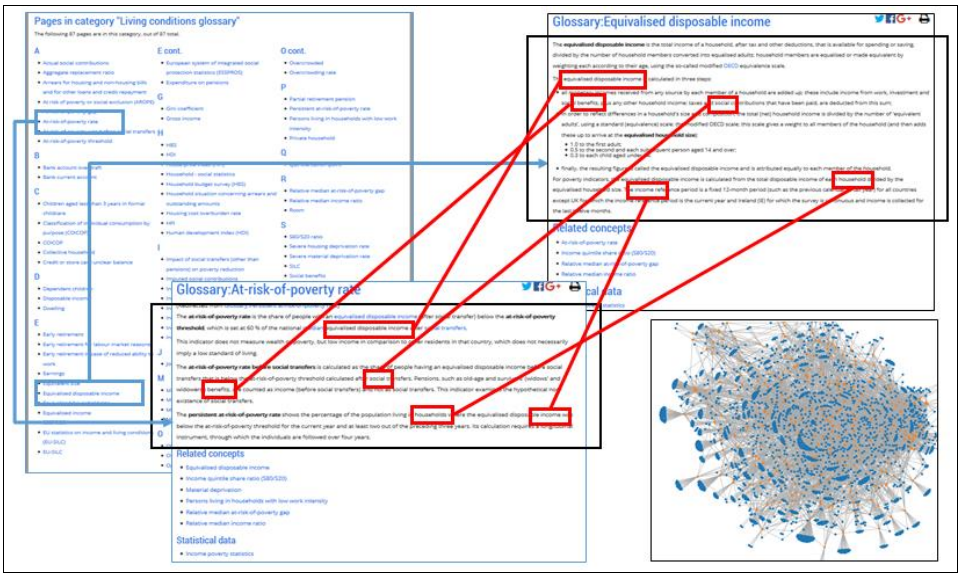

Figure A – Simple representation of the mapping process. Things (*e.g.*, concepts) appear related in the pseudo-KG (bottom right) through both the existence of (physical) weblinks (blue) and (statistical) co-occurences of other things (red) inside the textual description of these concepts.

4. CONCLUSION: STATUS AND WAY FORWARD

Because the approach adopted to build the KG is purely statistical and lacking advanced contextual analysis (the mapping is based on the co-occurrence of the terms analysed in the text documents), the search and discovery capabilities in the current protopytpe are showing strong limitations. We are therefore considering the introduction of Machine Learning (ML) techniques and Artificial Intelligence (AI) methods to improve (and enhance) the KG. Practically, we aim a training a word embedding model on the textual content of the *Glossary* and *Statistics Explained* to refine the entities and relations in the KG, *e.g.*, see what edges are likely to exist that do not currently exist, what edges are truly significant, what things are relevant, etc... In the future, this kind of automatic inference of relationships should allow us to further categorise inventory as it comes on the platform without requiring manual work. In parallel, we are also working on increasing the level of automation to enrich the KG with additional resources[12]. This will offer the possibility to deepen cross-domain and cross-policy knowledge, to combine insights from interdependent fields, with possibly highly domain-specific concepts.

[12] **https://github.com/eurostat/estatNet**

Once a critical mass of semantic information – including external resources – will have been integrated and processed, we can start thinking about making automatic inferences based on the things already in the KG while reasoning and defining abstract relationships. ML/AI can actually help to extend the KG (*e.g.*, through corpus-based ontology learning [7]), and in return, the KG can help to improve ML/AI algorithms (*e.g.*, through structural supervision [8]). This integrated approach ultimately leads to systems that will work like self-optimizing machines while being transparent to the underlying knowledge models.

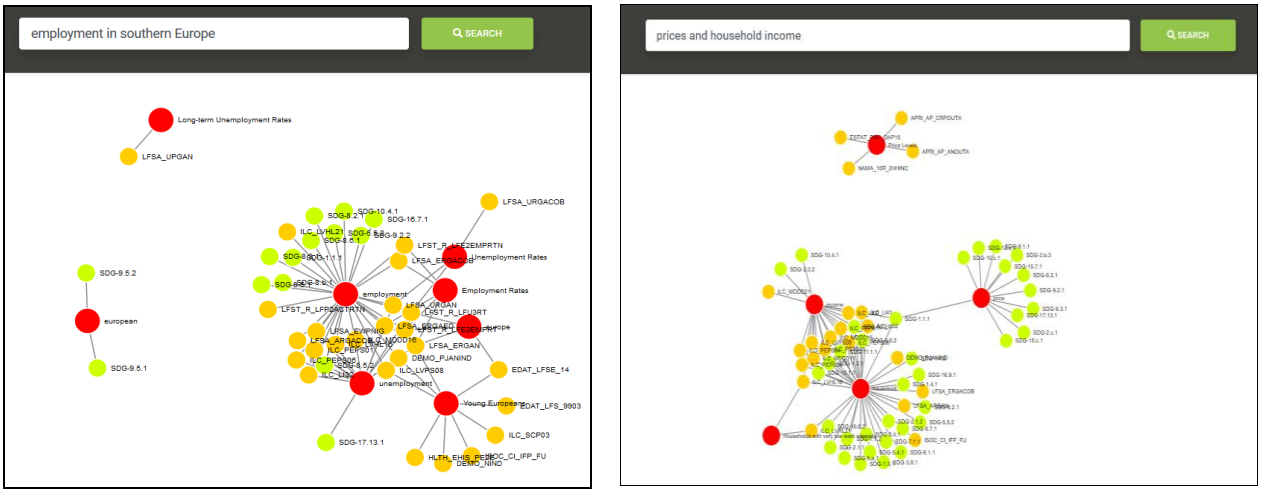

Figure B - Visual results of the *"employment in southern Europe"* (left) and "*prices and household income"* (right) queries over the pseudo knowledge graph. The links between related concepts (red nodes) and datasets (EU-SILC[13] and LFS[14] in orange, SDGs[15] in green) are represented.

REFERENCES

[1] E. Kalampokis *et al.* (2016): Open Statistics: The rise of a new era for Open Data?, doi: 10.1007/978-3-319-44421-5_3.

[2] E. Kalampokis *et al.* (2019): Interoperability conflicts in Linked Open Statistical Data, doi: 10.3390/info10080249.

[3] J. Hradec *et al.* (2019): Semantic Text Analysis tool: SeTA – Supporting analysts by applying advanced text mining techniques to large document collections, doi:10.2760/577814.

[4] S. Minaee et al. (2020): Deep learning based text classification: A comprehensive review, arXiv: 2004.03705.

[5] J. Shelke (2019): Topic classification using hybrid of unsupervised and supervised learning, San Jose State University.

[6] S. Ramnandan *et al.* (2015): Assigning semantic labels to data sources, doi: 10.1007/978-3-319-18818-8_25.

[7] G. Futia *et al.* (2020): SeMi: A SEmantic Modeling machIne to build Knowledge Graphs with graph neural networks, doi: 10.1016/j.softx.2020.100516.

[13] https://ec.europa.eu/eurostat/web/income-and-living-conditions/overview
[14] https://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey
[15] https://ec.europa.eu/eurostat/web/sdi

[8] T. Bui *et al.* (2018): Neural graph learning: training Neural Networks using graphs, doi: 10.1145/3159652.3159731.