Connecting Open Data and Sustainable Development Goals using a Semantic Knowledge Graph approach

José Eduardo Eguiguren[1], Nelson Piedra[1]

[1]Universidad Técnica Particular de Loja, Ecuador

{jeeguiguren, nopiedra}@utpl.edu.ec

Abstract. *This paper aims to present an initiative to establish links and relationships between Open Data (OD) and the Sustainable Development Goals (SDG). OD published by various organizations using the CKAN platform is highly dispersed and heterogeneous, making it harder to process and leverage those vast amounts of information in their current state. This paper approaches this opportunity by defining the steps that will allow the construction of a semantic representation of data sets found in public access data portals and each SDG using well-established ontologies and vocabularies. Doing so will provide the necessary tools to link the many resources containing data useful to the SDGs and relate them accordingly. Its implementation is a work in progress.*

1. Introduction

To achieve a better and more sustainable future for everyone, the United Nations General Assembly stablished 17 Sustainable Development Goals[1] with 169 targets to work towards. Their objective is to provide a solution to all the global challenges we face by 2030 (Hák, Janoušková, & Moldan, 2016). Amongst the many efforts that are being done to achieve each goal, capturing data to monitor and measure the SDG targets has become one of the most important (United Nations, n.d.). There are different initiatives that encourage the collection of data that supports SDGs. The Global Partnership for Sustainable Development Data was launched in 2015 in order to "strengthen data-driven decision-making" to help accomplish the SDG (Adams, 2015).

To fully implement and monitor progress on the SDG, decision makers, researchers, entrepreneurs and anyone interested need data and statistics that are open, reliable, accessible, accurate, timely, sufficiently disaggregated, relevant, and easy to use. Public access data portals are designed to collect and share open data. The publication of open data to which anyone access promotes transparency and collaboration. However, the primary obstacle presented by the nature of Open Data is the high level of diversity found between different datasets. Because standards for structuring data are vaguely used or often completely ignored, data portals end up with heterogeneous sets of data that further complicate its re-use and re-purpose. These portals can contain open data, at different levels of aggregation, related to the global SDG indicators, and that can be used to discover, understand, or communicate patterns and interrelationships between SDG and wealth of open data that are now available.

[1] See https://www.un.org/sustainabledevelopment/sustainable-development-goals/

Harnessing the power of digital transformation through open data and semantic technologies will be essential if we are to maximize the probability of implement and monitoring progress on the SDG. In this work, the authors propose a way to link open datasets with SDG indicators. The ultimate purpose is to develop the necessary tools for open data consumers to globally discover data related to the SDG, deploy new open data, and address these goals. By taking this opportunity, we can lay the foundation in creating the aforementioned processes, and draw a parallel between the datasets allocated in public access data portals and the SDG. Authors present a structure of Open Data retrieved from CKAN[2] repositories that are then represented semantically using the vocabularies DCAT[3] and Data cube[4]. The end goal is to link and relate them to the various indicators defined in the SDG. Section 2 describes some principles of Linked Data as well as concepts that will be used in this work. Then, section 3 defines the approach used in order to achieve said goals as well as some implementations that have been tested already. Finally, conclusions and future work are summarized in section 4.

2. Background

2.1. Linked Data

Tim Berners-Lee and his team in the World Wide Web Consortium (W3C) oversee the process of designing the Semantic Web (SW) that proposes the next step for the architecture of the Web. Its current goal is to provide a knowledge representation of LD as well as increasing the amount of Web resources that are easily interpreted by programs or agents (Alesso & Smith, 2006).

The Semantic Web defines the necessary practices and requirements to achieve a Structured Web. One of these is the use of Linked Data (LD) which is defined by (Bizer, 2009) as the "set of best practices to publish and connect structured data on the web". LD is also said to be the next step for Open Data where the information is interconnected, and can be related to other groups of information (Piedra, Chicaiza, López, & Caro, 2017).

LD's principles can be utilized in any field where the main objective is to centralize all information from various sources and stablish relationships that enrich the data that can be found. One of these examples, that satisfy the mentioned criteria, is the SmartLand-LD initiative (http://smartland.utpl.edu.ec) described in (Piedra & Suárez, 2018). One of its goals is to integrate every component that generates biologic, social, economic and environmental information distributed in different networks. By doing this they seek to provide a semantic structure in which its data will be exploited and linked to the Sustainable Development Goals.

The benefits of LD can be seen very clearly when doing a recollection of information across many different data sources. Such is the case for (Piedra, Tovar, Colomo-Palacios, Lopez-Vargas, & Alexandra Chicaiza, 2014) where LD principles are

[2] CKAN is a tool that easily allows the publishing and management of data sets. It is used globally by governments, companies and organizations. Using CKAN allows us to maintain a free access to every resource without restrictions, directly contributing to the main goal of Open Knowledge Foundation which is the organization behind the administration of CKAN.

[3] See https://www.w3.org/TR/vocab-dcat/

[4] See https://www.w3.org/TR/vocab-data-cube/

applied in order to enhance the discoverability of OpenCourseWare (OCW) distributed across many different universities. The solution provided includes the semantic integration of the information found in OCW to allow for better filtering options directed to end-users.

3. Proposed solution

The authors' solution arises from the following hypothesis: If there are so many Open Data resources found in many different platforms comprising a multitude of knowledge areas, there must be some of them that cover the same topics presented by each Sustainable Development Goal. The sheer volume of data may provide information useful to either track, measure or help each goal's progress. Instead of having all those sources of data separated in closed silos, we should integrate them and apply measures that help us gather useful information about them from a different perspective. That being the potential to link said resources to related SDG so that other people in need of such information can find it easily. By using multiple platforms from different countries and organizations we can cover a wide range of subjects that people from around the world can use in their efforts to achieve each goal established by the UN.

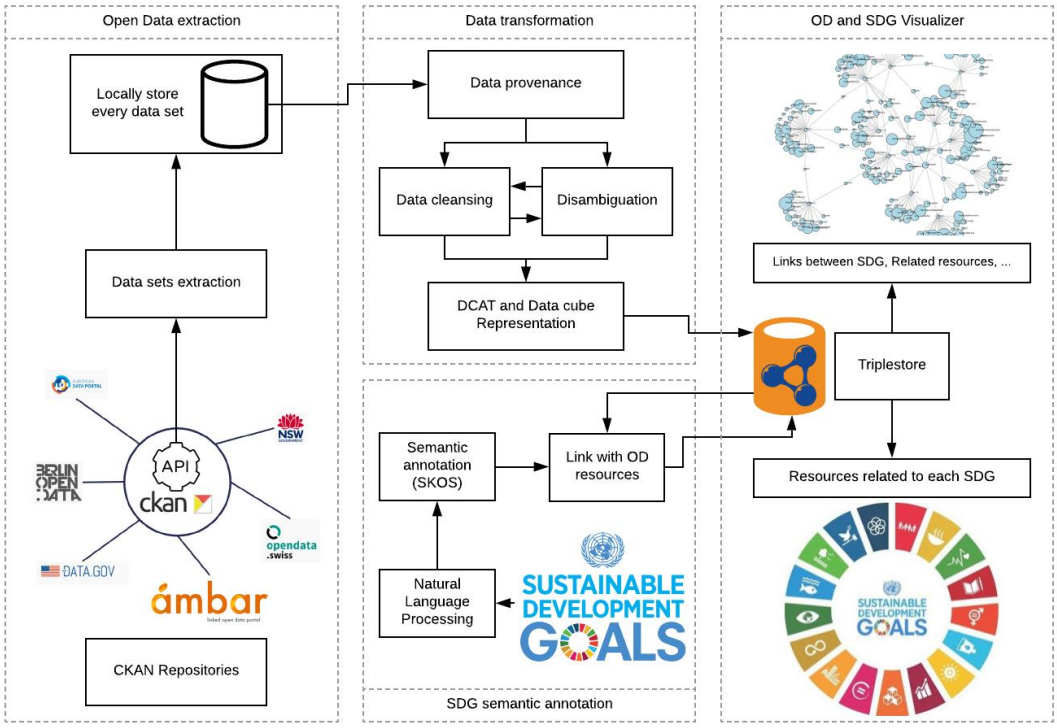

Figure 1. Architecture defined to build a semantic integration from public access data portals and link them with the SDG

In the next sections the authors describe the processes required to achieve our goal. These processes are automated and included in a web application. The application will then display the extracted and generated information in such a way that anyone can filter, view it, and use it according to his or her interests. This information includes the extracted data sets and the generated triples, as seen in Figure 1.

3.1. Extracting, transforming, linking and visualizing

The approach used in the process of building a semantic representation of OD found in CKAN platforms is divided in three stages. First, we extract datasets from various organizations that have implemented the CKAN platform to publish their data. Afterwards we transform all data found to homogenize it as much as possible. Then, continuing with the transformation, we use both DCAT and Data cube to build a semantic representation of the cleansed OD. Finally, we store the generated triples and proceed to the next phase.

The first challenge we encounter is finding a way to centralize all the available data sources into a single data silo. Thankfully, we can leverage the help that the CKAN ("CKAN," 2018) platform brings in order to standardize the means of publishing data and the extraction should be similar across the board.

Data visualization is one of the main end goals of this work where an infrastructure is built on top of the semantic information to provide solutions for end users to query OD, visualize and find similar datasets, integrate information to other sources such as the SDG, and much more.

3.2. Data cube representation

The reasoning behind using Data cube to represent statistical data is that within each resource, we may be able to find information that helps us describe the content of a data set. To achieve this transformation, we have automated the extraction process of any resource that has a CSV, XLS or XLSX file format. The values present in each file are then iterated and automatically annotated using the Data cube vocabulary. However, this process is currently experimental, and its results have not been used to relate the data sets to each SDG.

3.3. DCAT representation

The DCAT vocabulary provides the backbone of the semantic structure for public access data portals. Currently, after the data has been extracted automatically from the CKAN sources, the relational data is then transformed to match the vocabulary with the available metadata.

The transformation process has been automated using Apache Jenna[5] in Java. After all the data sets have been extracted and cleaned as seen in Figure 1, each one is mapped accordingly using the DCAT vocabulary automatically. The result is a set of triples that represent every extracted data set from each platform.

3.5 Transforming the SDG into a SKOS taxonomy

Once the data sets have been successfully transformed into a semantic knowledge representation, the authors need a way to find relationships between them and the SDG. Thus, the information present in each goal, target and indicator must be transformed into a semantic representation as well. The challenge faced by the authors is that the SDG are written in natural language. The following tool was found during research that tackles this exact problem.

[5] See https://jena.apache.org/

FRED is a tool whose purpose is to provide a machine reader for the Semantic Web. Alani et al., (2016) states how this tool uses multiple Natural Language Processing (NLP) components and unifies their result into a formal RDF/OWL graph. In this work FRED was used as a mediator in charge of analyzing the content of each goal, target and indicator from the SDG. The result provided by the tool is a formal representation of the most relevant knowledge of the SDG elements. FRED also has a key advantage that it automatically links the content found to already existing semantic knowledge such as DBpedia[6] resources.

The next step is to transform said graphs into an SKOS[7] representation so that we can stablish links with the resources from public data portals. The transformations process involves using a set of queries that can extract the elements shown in Figure 2. Each class is transformed into a SKOS concept that is then associated to a scheme based on the goal, target or indicator. The relationships found can be established using the related property of SKOS. If FRED found links to semantic web knowledge, we associate them with the concepts using the DC Terms property *subject*[8].

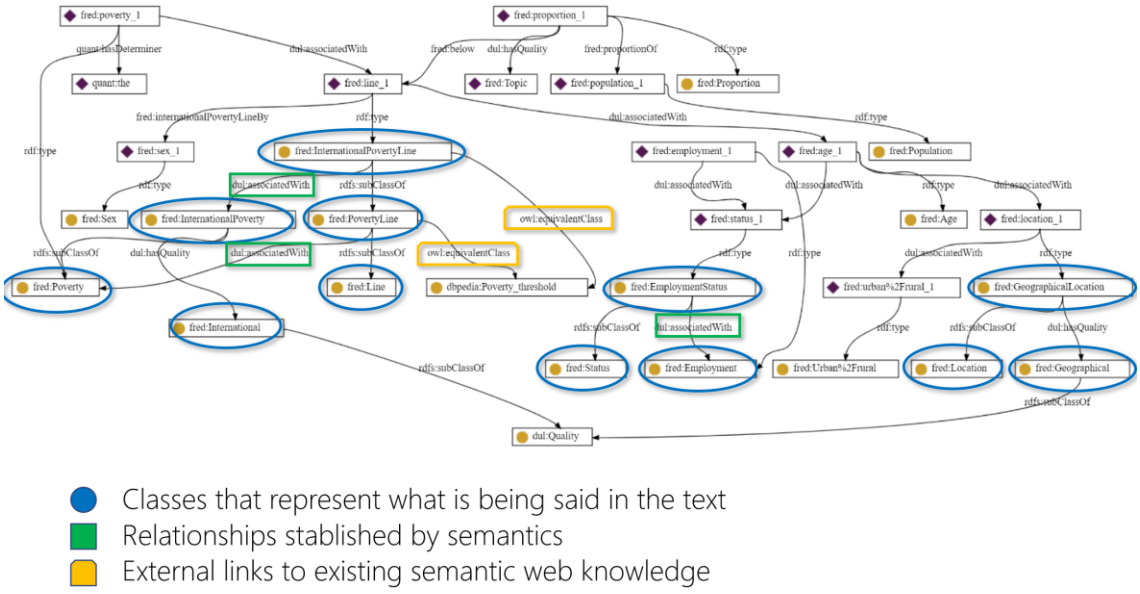

Figure 2. SDG Indicator 1.1.1 after being processed by FRED and the elements used in the SKOS transformation

3.5. Linking Open Data with the SDG

At this point the authors have successfully created a semantic knowledge graph representation out of Open Data and the SDGs. The final step was to find relationships between them that could tell us if a dataset has something to do with one or more SDG. This is done by analyzing the data found in every property from each dataset and finding similarities with the concepts from the SDG representation.

[6] See https://wiki.dbpedia.org/
[7] See https://www.w3.org/TR/swbp-skos-core-spec/
[8] See http://dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=terms#subject

The authors have chosen to use the SKOS vocabulary to represent the SDG for one main reason. The vocabularies used to represent OD already implement SKOS in their structure. This means that the authors can find similar concepts present in the SDG and OD representations. SDG representation the authors can discover which goals are data sets related to.

Our first approach involved analyzing the themes of each Data set and find matches with the concept taxonomies generated from each SDG. In this first iteration, we just analyzed the goals, excluding the targets and indicators to test the process with a smaller subset of information. We created a semantic property called "*automaticallyAnnotatedSubject"* that links any resource that has a match with an SDG. From 42,924 extracted data sets, we managed to create 6,295 links between them and the SDGs. Then the application is in charge of displaying which resources are related to each SDG. Useful information can be found using filtering tools provided by the app.

4. Conclusions and Future Work

In this work we have presented how we can leverage the advantages provided by the Semantic Web to allow for a better way to relate and distribute Open Data. A successful transformation from the current state of public access data portals to a more descriptive format using Linked Data can help overcome the numerous barriers that OD is currently presenting. We can also use the resulting representation to link related data to other fields such as the SDG. That way we can help to contribute in the global effort to achieve said goals in a very meaningful manner.

The challenges that were resolved during this work, can be organized into the following categories, a) technical issues: automatic extraction of open data from CKAN portals, identifying errors and data gaps, conversion to linked open data formats, maintaining Web interoperability standards and quality of the data, and creation of a web tool to managing these functionalities. b) Legal and provenance issues: datasets that lack explicit open licenses were assigned one based on the CKAN platform that contains them. c) Interpretation of multiple languages is a challenge that seeks to ensure no one is left behind; in this work, the prototype was applied only to datasets described in the English language. However, the authors consider as future work to meeting the needs of their users in multiple languages.

Future work includes finding more ways in which the Sustainable Development Goals can be related to Open Data. This process involves analyzing the semantic graphs that were produced using the previously discussed approaches. The authors have already managed to create an infrastructure in which relationships can be stablished between the SDG and OD. The following challenge involves the creation of useful information that details those connections by testing this process using all the data extracted from the previously mentioned open data portals.

Acknowledgement

The work has been funded and supported by the Universidad Técnica Particular de Loja (UTPL).

References

Adams, B. (2015). SDG Indicators and Data : Who collects ? Who reports ? Who benefts ? Policy Brief #9. *Global Policy Watch*, 1–8.

Alani, H., Gangemi, A., Presutti, V., Reforgiato Recupero, D., Giovanni Nuzzolese, A., Draicchio, F., & Mongiovì, M. (2016). Semantic Web Machine Reading with FRED. *Semantic Web*, *0*, 1–21. Retrieved from http://semantic-web-journal.org/system/files/swj1379.pdf

Alesso, H. P., & Smith, C. F. (2006). *Thinking on the Web: Berners-Lee, Godel and Turing*.

Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems*, (5), 87-92.

CKAN. (2018). Retrieved from https://ckan.org/

Hák, T., Janoušková, S., & Moldan, B. (2016). Sustainable Development Goals: A need for relevant indicators. *Ecological Indicators*, *60*, 565–573. https://doi.org/10.1016/j.ecolind.2015.08.003

Piedra, N., Chicaiza, J., López, J., & Caro, E. T. (2017). A rating system that open-data repositories must satisfy to be considered OER: Reusing open data resources in teaching. In *Global Engineering Education Conference (EDUCON), 2017 IEEE* (pp. 1768–1777).

Piedra, N., & Suárez, J. P. (2018). SmartLand-LD: A Linked data approach for integration of heterogeneous datasets to intelligent management of high biodiversity territories. *Advances in Intelligent Systems and Computing*, *688*, 207–218. https://doi.org/10.1007/978-3-319-69341-5_19

Piedra, N., Tovar, E., Colomo-Palacios, R., Lopez-Vargas, J., & Alexandra Chicaiza, J. (2014). Consuming and producing linked open data: the case of Opencourseware. *Program*, *48*(1), 16–40.

United Nations. (n.d.). United nations world data forum. Retrieved from https://undataforum.org/