PW 4: AI for Improved Data and Knowledge Governance

Alessio Nardin

September 11, 2023

Objectives

The primary objective of this assignment is to construct a knowledge graph (Hogan et al., 2021) of the Benetton Group's suppliers network. The secondary objective is to perform analytics and identify non-sustainable corporate behavior that impacts negatively human rights and the environment based on the knowledge contained in it. This specific theme was motivated by the announcement of the proposal Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937, which impacts these topics.

Motivation

From qualitative assessments (Villiers, 2022), the proposal Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937 has far-reaching implications for businesses operating within the EU, that increase the complexities of compliance, in particular concerning companies with global supply chains. Regulatory reporting obligations increase, in order to ensure the objectives of policy areas such as environmental regeneration and human rights protection of companies that operate economically in the European Union.

Benetton Group, like other companies operating in the EU, would be required to adhere to comprehensive due diligence processes concerning human rights and environmental impacts in its supply chain. In my opinion, the only sustainable option to address this challenge at the current state of technology is to leverage a combination of physical and digital infrastructure, enabled by semantic technologies.

For this reason, in this exercise we attempt to explore the problem and build a first prototype of a knowledge graph, starting from the publicly available supplier information of a company. The Benetton Group data was Findable, Accessible and Reusable. By inserting it into a knowledge graph, we attempt of making it Interoperable.

Methodology

The project is structured around the following sections:

1. Context

2. Data Cleaning
3. Reconciliation with OpenRefine
4. Serialization in RDF with rdflib
5. Querying the knowledge graph
6. Results

1 Context

The Benetton Group is an Italian textile company headquartered in Ponzano Veneto, in the Province of Treviso, Italy. It specializes in clothing production and operates a global network of franchised stores. These stores sell products under various brand names, including United Colors of Benetton, Undercolors of Benetton, and Sisley. It has a global supply chain containing 748 different suppliers of 25 different countries.

2 Data Cleaning

Data cleaning was performed with Python in the *DataCleaning.ipynb* notebook, mostly relying on the pandas library.

1. In the columns representing product types ('APPAREL,' 'ACCESSORIES,' and 'SHOES') the character 'l' is replaced with a Boolean value of True, while all other cells are filled with False.
2. Refinement of the 'ADDRESS' column includes identifying irregular entries by counting the occurrences of semicolons. These inconsistencies are manually corrected. The column is subsequently split into 'STREET,' 'CITY,' and 'ZIP' for more detailed data manipulation.
3. County information is segregated from both the 'ADDRESS' and 'ZIP' columns and compiled into a new 'COUNTY' column.
4. Identification and standardization of unique production phases are carried out. New Boolean columns represent these phases, and the data is populated based on these identifications.
5. Data pertaining to trade unions is standardized and refined. Extraneous spaces are removed, and missing values are populated with a default label of 'Unknown.'
6. Summary statistics are generated for the 'TRADE UNION' column to provide a snapshot of the distribution of True and False values.
7. Employee count information is restructured. The 'EMPLOYEES RANGE' column is split into 'MIN_EMPLOYEES' and 'MAX_EMPLOYEES,' and the original column is discarded to avoid redundancy.
8. The DataFrame is exported to a CSV file at several stages, serving as interim checkpoints for data validation. Additionally, the dataset is displayed for visual verification.

3 Reconciliation with OpenRefine

The alignment and validation of data were conducted using OpenRefine, mainly relying on Wikidata's reconciliation services. The goal was to synchronize geographical information columns with validated entity references from Wikidata.

The key steps in this process included:

1. **Main Alignment Operation**: The principal focus was on the `COUNTRY`, `COUNTY` and `CITY` columns. Wikidata's API was employed to match cell content with entities of type `country` (Q6256), `region`, and `city`..
2. **Entity Correlation and Auto-Matching**: OpenRefine used a 'standard-service' mode to align cell values with entities in Wikidata's identifier and schema spaces. Auto-matching was enabled to select the most accurate Wikidata entity when a strong match was detected.
3. **Alignment Engine Configuration**: The engine was set to function in 'row-based' mode, treating each row in the column independently and considering all rows as no facets were applied.
4. **Manual Reconciliation**: For each reconciled column, approximately 100 lines were reconciled manually as the service was not able to identify correctly the relevant Wikidata entry.

4 Transformation in RDF

We developed a command-line utility in Python to serialize the table in a Knowledge Graph. The utility is called by executing the *transform.py* file. The CSV file (downloaded from the OpenRefine project) named `data/csv/BenettonData.csv` serves as the input data source, while the turtle file `data/benettonsuppliers.ttl` contains the knowledge graph. In the following list we describe the components of the command-line utility.

1. ***initialize_graph()***: This function serves to initialize an empty RDF graph. It sets the base URI and namespaces.
2. ***initialize_classes()***: This function is responsible for populating the graph with instances of classes like `ProductType`, `Certificate`, and `ProcessType` in the ben namespace. It uses both a mapping dictionary and data from the CSV file.
3. ***initialize_properties()***: This function defines RDF properties such as `hasProcess`, `hasProductType`, and `hasCertificate`, which are used to delineate relationships between resources in the knowledge graph.
4. ***add_companies()***: The function iteratively reads rows from the CSV file, adding companies as instances of the `gr:BusinessEntity` class. It also sets various properties and values based on the data in the CSV file.
5. ***main()***: This function orchestrates the overall execution of the code. It reads the CSV data into a pandas DataFrame, initializes the RDF graph, and populates it with classes, properties, and company data. The graph is then serialized to a turtle file.

4.1 List of classes and properties

The outcome RDF (Resource Description Framework) graph generated by the provided Python code represents structured information about companies in Benetton's supply chain. This RDF graph uses various RDF classes, properties, and individual resources to capture different aspects of the supply chain data.

4.1.1 RDF Classes

- `ProductType`: Represents types of products in the Benetton supply chain.
- `Certificate`: Represents certifications that a supplier might have.
- `UnionInfo`: Contains information about trade unions associated with the supplier.
- `ProcessType`: Represents various processes carried out by the supplier's network.

4.1.2 RDF Properties

- `hasProcess`: Indicates the processes a company is involved in.
- `hasProductType`: Indicates the types of products a company produces.
- `hasCertificate`: Indicates the certificates a company holds.
- `minEmployees`: Indicates the minimum number of employees in the company.
- `maxEmployees`: Indicates the maximum number of employees in the company.
- `percentageOfMen`: Indicates the percentage of men in a company.
- `percentageOfWomen`: Indicates the percentage of women in a company.
- `percentageOfMigrants`: Indicates the percentage of migrant workers in a company.
- `hasWorkersRepresentative`: Indicates if an organization has workers' representatives.
- `hasTradeUnion`: Indicates if a company has a trade union.
- `tradeUnionName`: Name of the trade union within the company.
- `collectiveBargainingAgreement`: Indicates if a company has a collective bargaining agreement.
- `coverageCBA`: Percentage of the workforce covered by the collective bargaining agreement.

Certainly, let's expand on each of the queries and provide more context for each question:

"'latex

5 Querying the Knowledge Graph

Now that we have constructed a comprehensive knowledge graph, we can leverage it to extract valuable insights and answer specific questions related to the dataset. The Python code is in the *query.py* file. In this section, we will outline three example queries:

1. **Query 1: Percentage of Companies with Wet Process Manufacturing and Certificates**

   This query aims to calculate the percentage of companies within our dataset that engage in wet process manufacturing and possess certificates. It is essential to assess the prevalence of certifications among companies involved in wet process manufacturing, which is a particularly dangerous process for the health of the workers.

   Result: 45.2%

2. **Query 2: Employee Coverage by Collective Bargaining Agreements (CBA)**

   In this query, we seek to determine the average percentage of employees covered by Collective Bargaining Agreements (CBAs). We focus on companies that perform wet process manufacturing, produce apparel, hold certificates, and have established collective bargaining agreements.

   Result: 87.4 %

3. **Query 3: Companies Producing Shoes with High Female Workforce**

   The third query's objective is to compile a list of companies along with their names based on specific criteria. We will identify companies categorized as 'BusinessEntity,' ensuring they have names, are actively involved in 'Manufacturing' processes, particularly the production of 'Shoes,' and exhibit a substantial percentage of female employees, exceeding 70%. This query helps us identify companies promoting gender equality in the context of footwear manufacturing.

Table 1: Result of Query 3

Company URI	Company Name
`http://benettondata.it/graph/C122`	Albanian Shoes Corporation Nr 2
`http://benettondata.it/graph/C390`	Hangzhou Huayu Footwear Manufacturing Co., Ltd.
`http://benettondata.it/graph/C571`	Shandong Tuoda Shoes Co Ltd

By executing these queries on our knowledge graph, we can perform explorative analysis into various aspects of sustainability, labor rights, and gender equality within the companies represented in the dataset.

6 Results

This project produced several deliverables that support the reproducibility of the exercise. In particular:

- A cleaned and dataset of Benetton Supplier in CSV [1].
- A set of OpenRefine Instructions for reconciliation and further processing [2].
- A data transformation pipeline that integrates OpenRefine [3]

[1] Available here: https://doi.org/10.5281/zenodo.8336350
[2] Available here: https://doi.org/10.5281/zenodo.8336350
[3] The repository is available in https://github.com/AlessioNar/PW4

- A Knowledge Graph containing the supplier information of the Benetton Group in Turtle [4].
- A set of example SPARQL queries that aims to explore potential unsustainable corporate behavior [5]

All the following files and resources are published according to the FAIR data principles (Wilkinson et al., 2016; Nardin, 2023).

Conclusions

In summary, this exercise successfully achieved its objectives by constructing a knowledge graph of the Benetton Group's supplier network. The methodology involved data cleaning, reconciliation using OpenRefine, transformation into RDF, and the formulation of SPARQL queries.

The resulting deliverables include a cleaned dataset, OpenRefine reconciliation instructions, a data transformation pipeline, and a knowledge graph represented in Turtle format. Additionally, a set of illustrative SPARQL queries was developed to facilitate the investigation of human rights-related issues within the supply chain.

These resources adhere to the FAIR principles, ensuring their accessibility and reusability for further research and analysis. Ultimately, this exercise underscores the potential of knowledge graphs in enhancing data and knowledge governance, particularly in addressing sustainability challenges within complex supply chain networks.

References

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2021). Knowledge graphs. *ACM Comput. Surv.*, 54(4).

Nardin, A. (2023). Building a knowledge graph of the Benetton's group suppliers network.

Villiers, C. (2022). New directions in the european union's regulatory framework for corporate reporting, due diligence and accountability: The challenge of complexity. *European Journal of Risk Regulation*, 13:548 – 566.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. O. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S. C., Evelo, C. T. A., Finkers, R., González-Beltrán, A. N., Gray, A. J. G., Groth, P., Goble, C. A., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R. W. W., Kuhn, T., Kok, R. G., Kok, J. N., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R. C., Sansone, S.-A., Schultes, E. A., Sengstag, T., Slater, T., Strawn, G. O., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E. M., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and

[4] Available here: https://doi.org/10.5281/zenodo.8336350
[5] Accessible in the *query.py* file in the PW4 repository

Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.