Intensive Data Processing Systems - Assignment 1

Josep Maria Salvia Hornos - 48053698M

1 Introduction

The stars classification system proposed by Tim Berners-Lee is a widely recognized way of assessing the openness and linkability of data available on the web. The system assigns stars to data based on how open and accessible they are, with one star indicating basic openness and five stars indicating a high level of openness and linkability. The system is commonly referred to as the "TBL classification" and is widely used to assess the quality and accessibility of open data.

The objectives of this assignment are to explore and evaluate various sources of open data related to specific topics, including energy efficiency certificates, accommodation facilities, city council information, fuel consumption, and registration numbers in Spain. By researching and assessing these data sources, we aimed to understand how open they were based on Tim Berners-Lee's by stars classification system, which considers factors such as the availability of the data on the web, the data's structure, and the use of URIs to link to other data.

2 Energy Efficiency Lleida

Two sources:

Source 1: `https://certificacioenergetica.gencat.cat/icaen-visor/AppJava/services/certificats/show?language=es_ES`

The Lleida Energy Certificates dataset can be accessed through the website of the Generalitat de Catalunya. This dataset can only be searched through their filter platform, making it a **level 1** ★ dataset according to the TBL classification system.

The Energy Efficiency Lleida dataset is a 1-star dataset because it can only be accessed through a browser filter platform on the website of the Generalitat de Catalunya. The data cannot be accessed programmatically, which makes it difficult to automate any analysis or integrate it into other systems. The data is also not easily searchable or indexed, and the user must manually filter through the website to access the desired information. This lack of accessibility and usability greatly limits the potential use and value of the dataset.

Source 2: `https://analisi.transparenciacatalunya.cat/Energia/Certificats-d-efici-ncia-energ-tica-d-edificis/j6ii-t3w2`

It is available in a structured format, making it a **level 3** ★★★ dataset according to the TBL classification system.

The dataset is a rich source of information about the energy efficiency of buildings in Lleida. The dataset contains several important attributes, including the building's address, year of construction, non-renewable primary energy consumption, CO2 emissions, final energy consumption, and annual cost of energy per home, among many others. The information provided in this dataset can be used by policymakers, building owners, and energy experts to make more informed decisions on energy efficiency initiatives and policies. Overall, this dataset is rated as 3 stars due to its richness and relevance to energy efficiency.

3 Barcelona tourism

Link: `https://opendata-ajuntament.barcelona.cat/data/ca/dataset/equipament-centres-d-informacio`

The list of tourist accommodations in Barcelona can be found on the Open Data website of the Barcelona City Council. The dataset is available in a non-proprietary format and is structured, making it a **level 3** ★★★ dataset according to the TBL classification system.

The dataset provides a list of information centers located in the city of Barcelona. It contains a wide range of information about each center, including their register ID, name, institution ID, and institution name, road and neighborhood information among many others. Given the level of detail and completeness of this dataset, it can be classified as a level 3 dataset.

4 Open Data Paeria

4.1 Commercial Establishments that have Websites

List of Commercial Establishments in Lleida That Have Websites and Social Networks

Link: `https://aplicacionsweb.paeria.es/eOpenDataPublicWeb/faces/ca/cataleg/economia_i_empresa/ciencia-tecnologia/llista-comerc-smart-lleida/detalls`

The dataset is structured, available in non-proprietary format, making them **level 3** ★★★ datasets according to the TBL classification system.

The dataset is structured and available in CSV format, making it a level 3 dataset according to the TBL classification system. The header of the CSV includes important attributes such as the name of the commercial establishment, the type of association or neighborhood where it is located, the title or specialty of the establishment, the contact information (owner, telephone, email), the website and social networks (Facebook, Twitter, and other networks), and the Google Maps link. The dataset is useful for businesses, consumers, and researchers interested in the commercial activity of the city of Lleida and can be used for analysis and visualization of the data.

4.2 Population data

Population data classified by district, section, neighborhood, sex and by age groups on 1/1/2019

Link: `https://aplicacionsweb.paeria.es/eOpenDataPublicWeb/faces/ca/cataleg/territori/demografia/dades-poblacionals-2018/detalls`

The dataset is structured, available in non-proprietary format, making them **level 3** ★★★ datasets according to the TBL classification system.

The population data for Lleida, Spain, is available on the website of the city council. The data is classified by district, section, neighborhood, sex, and by age groups on 1/1/2019. The dataset is provided in a CSV format, making it easily machine-readable and reusable. The dataset can be used for various purposes such as research, urban planning, and public policy. With the granularity of the data, the dataset can provide insights into population patterns and demographic changes over time.

4.3 Map Of The Streets Of Lleida

Map Of The Streets Of The City Of Lleida And The Smaller Decentralized Entities

Link: `https://aplicacionsweb.paeria.es/eOpenDataPublicWeb/faces/ca/cataleg/territori/urbanisme_infraestructures/planol-carrerer-lleida/detalls`

The Map of the Streets of the City of Lleida and the Smaller Decentralized Entities dataset, available on the aplicacionsweb.paeria.es website, is a **level 1** ★ dataset according to the TBL classification system. The data is provided as a single PDF file, making it not machine-readable or indexable. Furthermore, there is no attached CSV, making it challenging to use the data in any automated processes. The lack of a structured format and the difficulty of extracting information from the PDF file make this dataset of limited use for data analysis or integration with other datasets.

4.4 Extra: News of the Technology Park

News of the Agricultural Science and Technology Park of Lleida

Link: `https://aplicacionsweb.paeria.es/eOpenDataPublicWeb/faces/ca/cataleg/economia_i_empresa/ciencia-tecnologia/noticies-parc-cientific-tecnologic-agroalimentari-lleida/detalls`

Unfortunately, the link provided for the dataset "News of the Agricultural Science and Technology Park of Lleida" is broken and leads to a 404 error. This means that the dataset is not accessible and cannot be used for any purpose. According to the TBL classification system, this dataset falls under the **-1 star category**. It is important to ensure that the links to datasets are valid and working, as broken links can cause inconvenience and disrupt the research process.

5 Fuel consumption in Spain

Link: https://datos.gob.es/es/catalogo/ea0010587-cantidad-total-y-cantidades-medias-consumidas-de-alimentos-bebidas-tabaco-combustibles-y-otras-fuentes-de-energia-epf-identificador-api-49141

The fuel consumption dataset for Spain is available on the datos.gob.es website. It is structured and available under an open license, making it a **level 3** ★★★ dataset according to the TBL classification system.

The dataset provides information on the total amount and average quantities consumed of various food items, beverages, tobacco, fuel, and other sources of energy in Spain. The data is categorized by expenditure codes with physical quantities and shows average quantities consumed over a period. It's worth noting that the dataset includes a mix of other consumables like rice, sugar, etc., and only shows data from families, not industry. The dataset also differentiates between diesel and gasoline fuel consumption, making it a valuable resource for researchers studying energy consumption patterns in Spain.

5.1 EXTRA: Well structured fuel consumption, La Rioja

Link: https://datos.gob.es/es/catalogo/a17002943-consumo-de-combustible1

The fuel consumption dataset for La Rioja is available on the datos.gob.es website. It is structured and available under an open license, making it a **level 3** ★★★ dataset according to the TBL classification system.

The fuel consumption dataset for "La Rioja" available on datos.gob.es website is a level 3 dataset according to the TBL classification system. This dataset is well-structured and available under an open license, providing easy access to anyone who wants to use it. The dataset contains information on fuel consumption in La Rioja in tons, measured monthly. The CSV file includes the header with the columns "[TIEMPO]", "[TIPO DE COMBUSTIBLE]", "[Measures].[Consumo en La Rioja (Toneladas)]" and "[Measures].[Var (%) mensual (La Rioja)]". One interesting aspect of this dataset is that it is specific to **La Rioja**.

Although this dataset has the same openness level as the previous one, this one is way richer in content and way more detailed.

6 Registration numbers Spain

Link: https://sedeapl.dgt.gob.es/WEB_IEST_CONSULTA/subcategoria.faces

The registration numbers dataset for Spain is available on the website of the Spanish Directorate General for Traffic. The dataset is structured, available in a non-proprietary format, and linked to related datasets, making it a **level 3** ★★★ dataset according to the TBL classification system.

It contains information on the type of vehicle, age, and registration numbers by autonomous community and province of residence. However, users must specify their query each time on the website (no canonical URL for all searches). Despite this inconvenience, the website provides well-structured information, making it an excellent resource for researchers interested in exploring transportation-related issues in Spain.

7 Extra: Data.gov

Link: `https://catalog.data.gov/dataset/lottery-powerball-winning-numbers-beginning-2010`

The Lottery Powerball Winning Numbers dataset is available on the data.gov website. It is structured and available in RDF format, making it a **level 4** ★★★★ dataset according to the TBL classification system.

The Lottery Powerball Winning Numbers dataset is a valuable resource for lottery enthusiasts and analysts. This dataset contains the winning numbers of Powerball lottery drawings in the US, beginning from the year 2010. It is available on data.gov and is structured in RDF format. The dataset is machine-readable and uses URIs to identify each resource, including each lottery drawing. Each item in the dataset contains information such as the draw date, winning numbers, and the multiplier. The RDF format also allows for easy linking to other datasets, providing more context for the analysis of the data. Overall, the Lottery Powerball Winning Numbers dataset is a well-structured and linkable dataset.

8 Conclusions

In conclusion, this assignment required researching various types of open data and evaluating their level of openness and linkability based on Tim Berners-Lee's stars classification system. Through this exercise, we were able to explore a variety of data sources related to energy efficiency certificates, accommodation facilities, city council information, fuel consumption, and registration numbers in Spain. Overall, we found that most of the data sources were at least level 3 in terms of openness and linkability, indicating that they were available in a non-proprietary format and had URIs to point to other data, although some sources had limitations such as non-UTF-8 encoding or poor website design. By understanding the level of openness of different types of data, we can better utilize open data for research, analysis, and decision-making purposes.