Assignment 1 – Animal Scouter

University of Aveiro

Web Semântica

2022/2023

Group nº 2:

* João Bernardo Coelho Leite
* João Pedro dos Reis
* Luís Miguel Gomes Batista



Index

[Introduction 2](#_Toc130768228)

[Data, sources and transformations 3](#_Toc130768229)

[Dataset 3](#_Toc130768230)

[Selection of serialisation format 3](#_Toc130768231)

[Transformations 3](#_Toc130768232)

[Overview 3](#_Toc130768233)

[References 5](#_Toc130768234)

Introduction

The Semantic Web is an extension of the World Wide Web that allows data to be shared and reused across different applications, platforms, and systems. It is based on the principle of creating machine-readable data that can be easily understood and processed by computers. One of the key components of the Semantic Web is the use of linked data, which refers to a set of best practices for publishing and connecting structured data on the web.

In this report, we will discuss the conversion of a database from CSV (Comma Separated Variables) to N-Triples, a *Triplestore* database format, using a Python script. We will also explore the use of SPARQL, a query language used to retrieve and manipulate data stored in RDF formats. Additionally, we will showcase the features of a website that has been created to demonstrate the capabilities of the Semantic Web, including the ability to query and retrieve data using SPARQL.

The report will begin by describing the data and process of converting a database from CSV to N-Triples, and the advantages of using a *Triplestore* database. It will also be explained the basics of SPARQL and how it can be used to query and retrieve data from *Triplestore* databases.

Finally, the website that has been created to showcase the features of the Semantic Web will be discussed. The website will include a demonstration of how to use it, as well as examples of how the data can be visualized and analysed.

Overall, this report aims to provide an introduction to the Semantic Web, and to demonstrate the benefits of using linked data and *Triplestore* databases for managing and sharing structured data on the web.

Data, sources and transformations

## Dataset

The Zoo Animals dataset used in this Semantic Web analysis was sourced from Kaggle, a popular platform for data science and machine learning projects. The dataset consists of two separate CSV files, each containing 18 columns of data related to various attributes of different zoo animals.

### Selection of serialisation format

Before any transformations there was an important decision to make regarding what type of data should be used NT, N3 or RDF/XML. It ultimately was decided that NT (N-Triples) was the right choice for this project. The NT format is a widely accepted and standardized format for representing data in Semantic Web, it is designed to represent structured data providing ease of querying and analysis. Lastly, it’s very optimised for scalability being able to handle large amounts of data efficiently.

### Transformations

In order to prepare the dataset for use in the Semantic Web, several transformations were applied to the data. Firstly, the two CSV files were merged into a single file to simplify the data and make it easier to work with. Next, a script was created to automatically convert the merged dataset into the NT (N-Triples) format, which is a standard format for representing data in the Semantic Web. The script to generates unique IDs for the "class type", "animal name", and "type of nurturing" attributes in order to be able to name them separately to one another. These IDs served to create triples consisting of a subject, predicate, and object as per the example:

<http://zoo.org/animal/id/turtle> <http://zoo.org/pred/name> "Turtle" .

Example 1 Triple where a subject "id" has “name" which is "Turtle"

## Overview

After the data was converted to the NT format, it was loaded into a *Triplestore*, which is a type of database designed for storing RDF data. This allowed for efficient querying using SPARQL and analysis of the data, as well as integration with other Semantic Web tools and technologies such as GraphDB.

Overall, the use of the Zoo Animals dataset allowed for a robust and interesting analysis of the data using Semantic Web techniques. By applying transformations to the data and converting it into the NT format, it was possible to unlock valuable insights and information about the various attributes of different zoo animals.

References