Knowledge Graph of Press Clippings Referring Social Minorities*

Paulo Martins[1 [0000−0002−1521−0014]], Leandro Costa[2 [0000−0003−4973−1093]], and José Carlos Ramalho[3 [0000−0002−8574−1574]]

[1] University of Minho, Portugal paulo.jorge.pm@gmail.com
http://www.paulojorgepm.net
[2] University of Minho, Portugal leandro.costa16@hotmail.com
[3] Department of Informatics, University of Minho, Portugal jcr@di.uminho.pt
https://www.di.uminho.pt/~jcr

Abstract. *Major Minors* is a project that collects press clippings from Portuguese newspapers (currently from 1996 until 2019) which refer subjects related with minorities. Its datasource is the Arquivo.pt (repository of the past Portuguese World Wide Web). This data was used to generate ontologies (RDF triplestores composing a semantic database) and interfaces to interact with them (SPARQL APIs following W3C standards for Semantic Web). We enriched this basis with new ramifications, by identifying and crossing references with 19 entities. This paper describes the methodologies implemented to develop this Knowledge Graph.

Keywords: Ontology · Knowledge Graph · RDF · OWL · Minorities.

1 Introduction - *Major Minors*

The stigma associated with certain minorities is continuously changing, yet there's a lack of central data repositories for tracking and studying this representation. Media, specifically published articles on renowned digital newspapers, can be used as a source for analyzing this[3], being it through the journalistic representation (text and photo illustrations) or user expression (comments).

This article focus a project developed with the objective of extracting and giving meaning to Big Data related with this subject, making it available as a research tool. It is entitled *Major Minors* and was submitted to the *Arquivo.pt* 2021 annual competition, promoted by FCT, winning the 1st prize. It uses the *Arquivo.pt* (repository of the past Portuguese World Wide Web) as its main data source, crawling newspapers from the first two decades of the XXI century.

Major Minors is therefore a project aimed at identifying and collecting, from digital newspapers, press clippings that refer subjects related with social minorities, building ontologies centered around articles, images and user comments. In a second stage, these data were crossed with dozens of datasets, identifying entities referred in the texts (e.g. Political Parties, Persons, Brands, etc.), generating a Knowledge Graph of these relationships and publicly available interfaces and APIs for navigating them. The corpus is treated and stored using technologies

recommended by W3C for the Semantic Web, mainly RDF (Resource Description Framework), OWL (Ontology Web Language), Turtle (Terse RDF Triple Language) and SPARQL (Sparql Protocol and RDF Query Language).

Ontologies, the backbone of the Web 3.0, which contain the vocabulary, semantic relationships, and simple rules of inference and logic for a specific domain, are accessed by software agents. These agents locate and combine data from many sources to deliver meaningful information to the user[7].

This project explores this principle, trying to contribute to the Semantic Web initiative, by building a web of relational semantic data related to a domain (Web 3.0), instead of a traditional web of documents (Web 2.0). Our subject of study is the media representation of minorities in Portuguese newspapers, mainly the "Público", because it is the oldest daily newspaper archived in the "Arquivo.pt", with the largest number of articles available. At the moment, this project archived ∼49.000 articles, in ontological graphs, between 1996 and 2019. These ontologies were augmented into a Knowledge Graph, by extracting and building relationships with real-world entities mentioned in the texts.

For this study, 8 minority groups were focused: refugees, women, Africans, Asians, homosexuals, migrants, gypsies and animals. By "minorities" we refer to macro social groups with some kind of social stigma and/or struggle for equality/rights. The chosen categorization reflect the research fields of the partnerships established with other research groups (CEHUM, NetLang, etc.).

This project was born in he Department of Informatics of the University of Minho, but we underline the multidisciplinar collaborations later established with other Research Centers, namely humanities research groups (e.g. CEHUM): the chosen categorizations reflect the fields of study of these collaborations partnered with the project (the most recent one is the NetLang R&D project). Research is undergoing and new minorities and thematics could be integrated in the future, accordingly to new partnerships.

This paper objective is to focus the methodologies developed for the Knowledge Graph implementation. In a past paper[6] we covered an overall overview of the project, for that reason and space limitations, we will be succinct and avoid other subjects. On section 2 we'll briefly summarize an overview of the main development stages for contextualization. On section 3 we'll cover the methodologies related with the ontology. Finally, section 4 will summarize these results.

2 Project Overview

From this project resulted a website with different services, two ontologies (ontology *a)* all the newspaper corpus; ontology *b)* only the corpus referring minorities), public APIs (SPARQL based) and Reactive Interfaces to mediate data exploration. Namely, these are some of the main endpoints/URLs:

- ***Major Minors* website:** http://minors.ilch.uminho.pt
- **RDF Triplestore (GraphDB):** http://sparql.ilch.uminho.pt
- **Open-source tools (GitHub):** https://github.com/Paulo-Jorge-PM

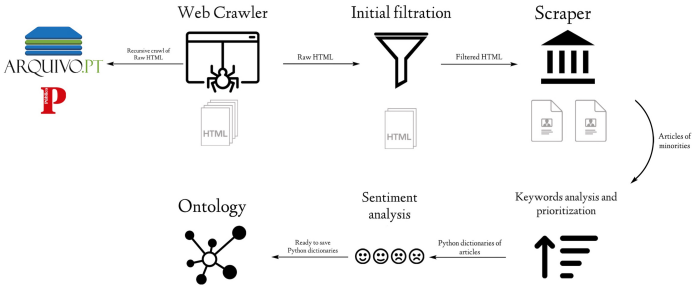

Fig. 1. Roadmap of the project main development stages.

This project development cycle is summarized on Figure 1.

The initial stages extracted and segregated data from the raw newspaper articles, prioritizing them accordingly to the identification of semantic trees of keywords related with specific minorities, and their position/relevance inside the texts of each article. For that we built an algorithm in which different scores were attributed to different types of keywords and positions (e.g. if mentioned on the title, body, description, tags, etc.), allowing us to order the articles, comments and photos by their relevance relative to each identified minority (Figure 2 exemplifies this prioritization algorithm). The objective of these initial processes was to identify articles mentioning subjects associated with different minorities, ordering by its relevance, and to extract structured metadata.

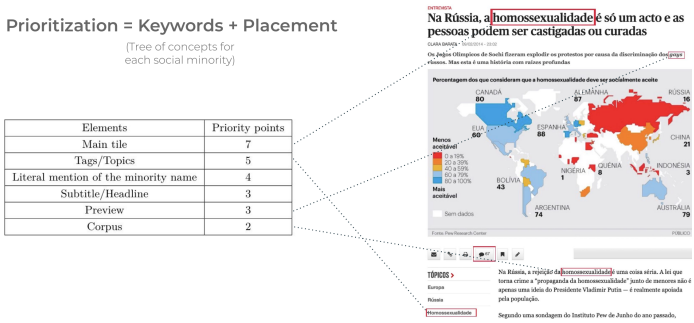

Fig. 2. Example of the prioritization algorithm score system.

The last and central stages of development focused the generation of ontologies with these metadata, and the augmentation of the graph three with new ramifications by identifying, inside each article, references to external entities (Public Figures, Political Parties, Cities, etc.), building a Knowledge Graph with

contextual relationships. Different interfaces and APIs were built around these data, in order to make them publicly accessible and easy to navigate (Figure 3).

3 From Ontologies to a Knowledge Graph

The central stages of this project focused the development of ontologies, tools for automatically feeding them into a Knowledge Graph, and interfaces for data availability. Figure 3 summarizes this approach.

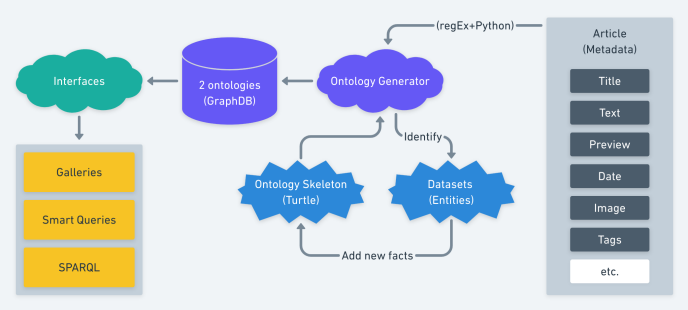

Fig. 3. Overview of the ontology generation.

An ontology is a "specification of a conceptualization", "a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents"[4]. Essentially, in a broad sense, it is a type of graph database based on triplestores for each fact (Subject, Predicate, Object), built around a semantic representation of the relationships between them. Guarino et al. (2009)[5] expanded on what is a conceptualization and the history of the formal definition of ontologies, suggesting more precise ones.

Knowledge Graphs are clusters of ontologies with common ramifications and ways to efficiently interact with them. They are important because "are critical to many enterprises: they provide the structured data and factual knowledge that drive many products and make them more intelligent and magical"[8].

Google popularized the concept of Knowledge Graph with the launch of their project in 2012, aimed at improving their search engine feedback[10]. Since then, many projects have been trying to translate Big Data into comprehensive data through similar methodologies, for example the ones related with the Linked Open Data initiative, WordNet, YAGO[9], DBPedia, etc. These approaches are expanding the concept of what we understand to be the World Wide Web, building the basis of the Web 3.0, also known as the Semantic Web.

This project gets inspiration from the mentioned ones, implementing personalized methodologies based upon entities identification: a graph of articles and metadata, expanded with relationships of entities identified inside them.

These approaches reduce the man-machine gap, by giving semantic and structural meaning to complex Big Data. SPARQL, RDF, and OWL, expressed through Turtle, were central technologies for this stage, both recommended by W3C for implementing Web Ontologies, trying to revolutionize the way we interact with the Web, data and computers. Because of these characteristics, ontologies are one of the key technologies for the implementation of the new generation of the World Wide Web.

3.1 Ontology Generator

The ontologies underlying this project were generated dynamically. Initially we built a skeleton of the ontologies using the software *Protègé*, generating a schema with the main static structures, but not specific individual data. We saved the ontology in the *Turtle* syntax, always working with this format as a basis.

Later we built a tools for automatically identifying metadata from the articles and entities inside the texts, dynamically expanding the ontology graph by adding new contextualized *Turtle* segments to the initial ontology, expanding it with individuals and ramifications, adding new facts to the pre-defined skeleton.

In order to identify external entities mentioned in the articles, an intermediary step was taken, where we developed datasets with thousands of pre-defined real-world entities, divided into 19 categories. Some entities are complete (e.g. Countries, Capitals, Continents, Sports, Months, etc.), but others are incomplete because of their nature, a constant work in progress (Public Figures, Brands, Political Parties, etc.). Various strategies were adopted in order to collect this datasets (crawling public figures from magazines and newspapers; sources like Wikipedia or "dados.gov"; intensive use of RegEx etc.). This work was contextualized to each one of the 19 datasets developed. This article, due to space limitations, does not intend to focus the development of the datasets, but this brief contextualization was important to understand how the ontology generator operates. For example, at the moment, we identified in the articles references to 32.648 Persons, 8.525 Political Parties, 3.799 TV Channels, 28.933 Cities, etc.

The ontology generator uses these pre-defined datasets with external entities, identifying them inside the article using Regular Expressions - if it identifies one entity, it generates a new Turtle segment adding new contextual graph ramifications. This approach transforms the ontology into a Knowledge Graph, because it is not anymore a singular a ontology of press clippings, it is a cluster of relationships between them and external contextual data with a common interface.

3.2 Interfaces for data availability

The rich ramifications of entities integrated into this Knowledge Graph, allows the end-user to define intuitive SPARQL queries to extract very complex data, e.g.: extraction of articles from a set of years, referring a certain Public Figure, without a particular job, referring a specific Political Party and Brand, in may, with a set of keywords related to two minorities, etc. *ad infinitum*.

We built different interfaces and layers in order to interact with this data through SPARQL queries (W3C recommended technology for querying Web Ontologies) and APIs. Even though SPARQL is very accessible and easy to learn, usually this technology is more commonly used by a specialized public. For this reason, since we wanted to make this project available to the general community (academics from all areas, journalists, educators, hobbyists, etc.), we opted to build 3 different layers of access to the Knowledge Graph:

- A static interface[4], exposing 3 galleries (articles, images and comments) which any kind of user can navigate, but the interaction is very limited;
- A second layer for intermediate users[5], with a Reactive Interface that interacts with a visual form with 3 levels of filtering (main classes, children, and some keywords), building automatically, in real-time, SPARQL queries that communicate with the Knowledge Graph, extracting richer data. This interface is open-source and can be easily adapted to any kind of ontology;
- A third layer[6], for specialized public, with 2 endpoints, giving direct access to the triplestore database and graph interfaces, and APIs to execute SPARQL queries, with unlimited freedom for manipulating the Knowledge Graph.

This project main objective is to facilitate the research of minority questions, by providing easy and intuitive access to a contextual Knowledge Graph of complex Big Data, usually inaccessible from the general nontechnical public. In the background, we opted for GraphDB to expose these triplestores.

3.3 Choice of triplestore databases: pros and cons of GraphDB

This project worked with the Turtle syntax to generate the ontology (alternative RDF syntax). For making the data publicly available we opted for a dedicated database triplestore to store it and SPARQL compliant APIs for querying it.

We analysed and considered 11 dedicated ontological triplestore databases: GraphDB, StarDog, AllegroGraph, AnzoGraph, BlazeGraph, 4store, Apache Jena, Virtuoso, MarkLogic, RDFox and TerminusDB. We aimed for open-source or limited free versions (for that reason some were immediately discarded because they had proprietary licenses with fees, e.g. RDFox), SPARQL native full support (for that reason e.g. TerminusDB was rejected) and, secondarily, provided public beautiful visual interfaces for facilitating data exploration (for that reason, AnzoGraph, for instance, was rejected in later stages of selection).

We started by analysing published benchmarks and articles. According to these data and initial criteria, our main preferences were GraphDB, StarDog and AnzoGraph. Both have top performance, intuitive APIs and interfaces for visually interacting with the ontologies. Mainly: both offer free versions.

The majority had proprietary licenses, but some had also a free alternative limited plan, for example: GraphDB free plan limited it to 2 simultaneous

[4] Galleries: http://minors.ilch.uminho.pt/articles
[5] Reactive Interface: http://minors.ilch.uminho.pt/search
[6] DB: http://sparql.ilch.uminho.pt; API: http://minors.ilch.uminho.pt/sparql

queries, AnzoGraph had a limit of 8GG RAM use (increased to 16GB with a annual registration), and StarDog (the best hybrid license).

The 3 final choices were a balance between free, performative and interfaces. AnzoGraph, even though where the most performative in some benchmarks[1], had an inferior visual interface, for that reason were rejected. In the end it was a tie between GraphDB and StarDog. We ended opting for the first one, because in general benchmarks were better for specific queries. For Belini et al. (2018) "Virtuoso performs better in presence of less selective queries(...) on the contrary, GraphDB performs better when specific results are searched"[2], but also notes that Virtuoso has inconsistencies constrains.

Even though we opted for the GraphDB Free Edition, the limit of 2 simultaneously queries ended up to be a bottleneck and we ended up developing a cache system for improving that limitation. For future projects wit similar context we would advise for StarDog instead. Otherwise, the benchmarks differences are not that impactful between the major contenders, all have different particularities, the context of each project should be considered.

4 Conclusions

Major Minors provides an open-source Knowledge Graph of press clippings focused on social minorities. It expands ontologies with semantic relationships between thousands of real-world entities, augmenting data manipulation.

Various projects, e.g. YAGO[9], apply similar techniques, but are more generic. This project tried to developed its own methodology to achieve similar results focusing on a singular thematic (minorities). The applied approach started by developing isolated datasets of entities, created individually through crawlers and Regular Expressions, of real-world entities, that feed an ontology generator built to transform an ontology skeleton into an augmented Knowledge Graph. In total, the datasets had 16.096 individual entities, divided in 19 categories. Algorithms for prioritization and sorting were developed alongside the entities identification. This article summarized those methodologies.

The resulting data were stored using triplesotres databases, exposing them through SPARQL queries. Accordingly to our context and past benchmarks, we concluded that GraphDB and StarDog were the better choices for our necessities. Currently, this project has 5 million triplestores, 49.000 articles archived and 8 macro minority groups with references to 256.458 entities.

Various Research Groups partnerships have been established (CEHUM, NetLang, etc.) besides the original ones (DI) in order to give longevity to this open-source project through applied studies. In future iterations and partnerships, we would like to expand the corpus to others timelines, newspapers and thematics. We invite the research community to contribute to this expansion.

References

1. Addlesee, A.: Comparison of Linked Data Triplestores: Developing the Methodology. Medium (May 2019), https://medium.com/wallscope/comparison-of-linked-data-triplestores-developing-the-methodology-e87771cb3011
2. Bellini, P., Nesi, P.: Performance assessment of rdf graph databases for smart city services. Journal of Visual Languages & Computing **45**, 24–38 (2018). https://doi.org/https://doi.org/10.1016/j.jvlc.2018.03.002, https://www.sciencedirect.com/science/article/pii/S1045926X1730246X
3. Bleich, E., Stonebraker, H., Nisar, H., Abdelhamid, R.: Media portrayals of minorities: Muslims in british newspaper headlines, 2001–2012. Journal of Ethnic and Migration Studies **41**(6), 942–962 (2015)
4. Gruber, T.: Ontology. Encyclopedia of database systems **1**, 1963–1965 (2009)
5. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Handbook on ontologies, pp. 1–17. Springer (2009)
6. Martins, P.J.P., Costa, L.J.A.D., Ramalho, J.C.: Major Minors - Ontological Representation of Minorities by Newspapers. In: Queirós, R., Pinto, M., Simões, A., Portela, F., Pereira, M.J.a. (eds.) 10th Symposium on Languages, Applications and Technologies (SLATE 2021). Open Access Series in Informatics (OASIcs), vol. 94, pp. 3:1–3:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2021). https://doi.org/10.4230/OASIcs.SLATE.2021.3, https://drops.dagstuhl.de/opus/volltexte/2021/14420
7. Morris, R.D.: Web 3.0: Implications for online learning. TechTrends **55**(1), 42–46 (2011)
8. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., Taylor, J.: Industry-scale knowledge graphs: lessons and challenges: five diverse technology companies show how it's done. Queue **17**(2), 48–75 (2019)
9. Pellissier-Tanon, T., Weikum, G., Suchanek, F.: Yago 4: A reason-able knowledge base. ESWC 2020 (2020), http://yago-knowledge.org
10. Singhal, A.: Introducing the knowledge graph: things, not strings. Official google blog **5**, 16 (2012)