Prémio Arquivo.pt

Descrição Sumária do Trabalho

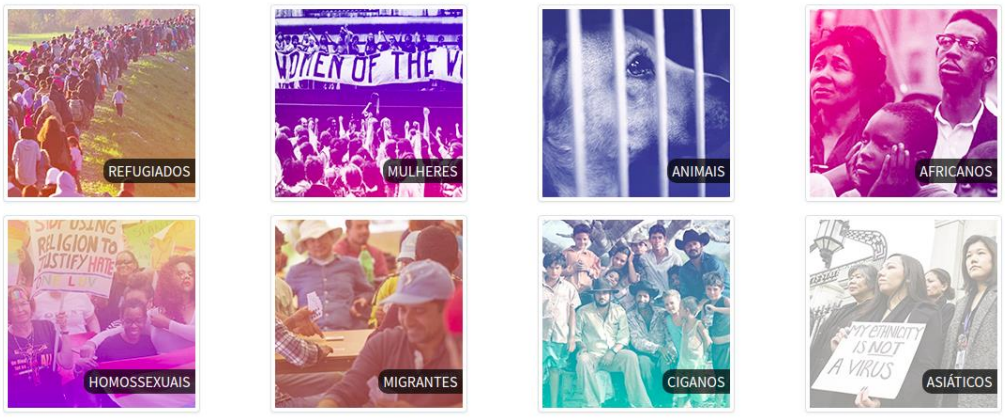
Identificação

- **Título:** *Major Minors* – Representação de minorias pelos jornais portugueses: uma base de dados ontológica.
- **Área temática:** Informática; Humanidades; Ciências Sociais; Linguística Computacional.
- **Candidatos:** Paulo Jorge Pereira Martins / Leandro José Abreu Dias Costa.
- **Email:** paulo.jorge.pm@ilch.uminho.pt / leandro.costa16@hotmail.com
- **Website:** http://minors.ilch.uminho.pt

Descrição do Trabalho

O presente trabalho consiste na criação da primeira base de dados ontológica portuguesa (*Web Semântica 3.0 / RDF*), que visa mapear e estudar a representação de minorias em contexto jornalístico português ao longo das duas primeiras décadas do século XXI. O *corpus* inclui uma base de dados semântica, acessível mediante *queries SPARQL*, dividida em três tipos de recortes de imprensa: artigos, comentários de utilizadores e imagens. O projeto nasce de dois estudantes e investigadores de pós-graduação da Universidade do Minho, contando com a colaboração de Grupos de Investigação em Humanidades e Engenharia Informática. Todo o material resultante do projeto está disponibilizado em acesso aberto através do seu *website* (licença MIT).

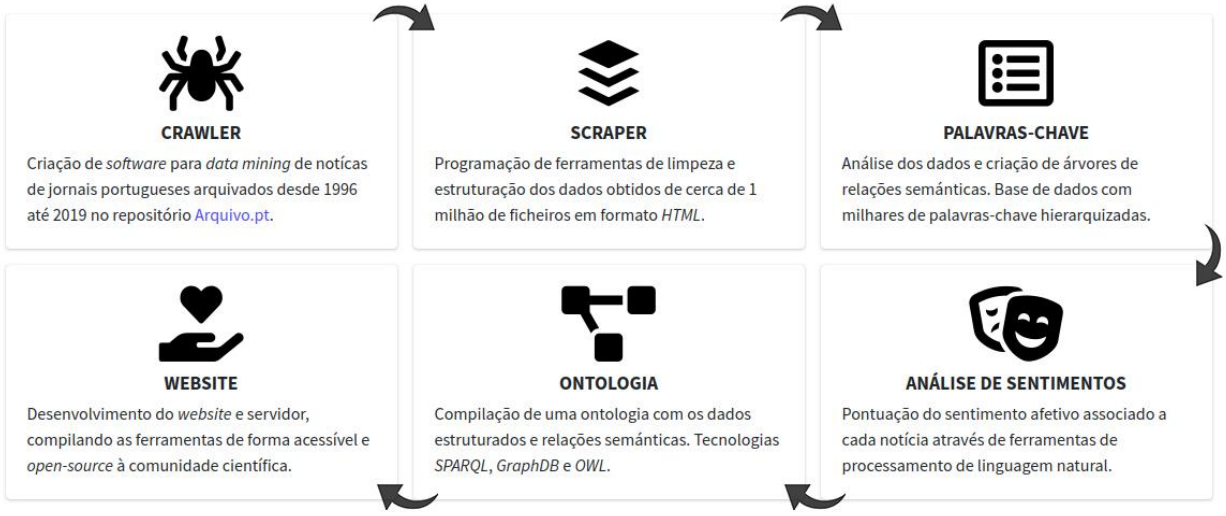
As minorias em estudo agrupam-se da seguinte forma:



Este acervo de dados pode ser definido como um mapa sociocultural que permite traçar a imagem pública e representação social tanto de minorias de longa data (ligadas a questões raciais), como de minorias emergentes (a contemporaneidade dos refugiados).

Por minorias entenda-se grupos que sofrem de algum tipo de discriminação. O leque de grupos e possíveis abordagens seriam extensos, mesmo a própria definição de minoria é desafiante. As 8 minorias selecionadas para este estudo coincidem com os vetores de investigação do Centro de Estudos Humanísticos e do Departamento de Informática da Universidade do Minho. A base de dados e ferramentas (crawler, scrapper e construtores da ontologia) são modulares, *open-source* e facilmente expandíeis para novos critérios de extração. Os modelos e estruturas adotados refletem também os interesses e necessidades dos Grupos de Investigação associados, com vista a realizar trabalho de pesquisa futuro no contexto de *Natural Language Processing* (comentários dos utilizadores e *corpus* jornalístico), Inteligência Artifical e Reconhecimento de Imagem (fotografias ilustrativas de cada artigo), etc.

A componente técnica deste projeto pode ser descrita em seis fases distintas:



Nas fases iniciais de *data mining* (recolha e tratamento de dados em bruto), a plataforma de referência para extração dos dados foi o *Arquivo.pt*. Através dela, até ao momento extraímos mais de 1 milhão de ficheiros em formato *HTML*, dos quais cerca de 40 mil apresentavam referências e palavras-chave de interesse para o domínio em estudo. Numa primeira fase o jornal selecionado para estudo foi o *Público*. Esta opção deve-se a este ser um dos jornais em formato digital mais antigos e bem reputados em Portugal (o repositório *Arquivo.pt* contém as primeiras edições de 1996. Se o interesse da comunidade assim o ditar expandiremos o projeto com novos *corpus*.

A partir da recolha de centenas de milhares de artigos presentes no *Arquivo.pt*, foram extraídos e filtrados dados de interesse para construir a presente árvore semântica em estudo. Foram criadas árvores de relações (*grafos*) e referências cruzadas entre os diferentes recortes de imprensa e milhares de palavras-chave, grupos de entidades (pessoas, animais, lugares, religiões, partidos políticos, etc.). Estas relações com semântica visam permitir dar a conhecer, de forma mais aprofundada, a representação das minorias nos *media* portugueses, servindo de material de estudo rico para obtenção de dados não possíveis em bases de dados tradicionais.

Com os artigos recolhidos foi realizada uma filtração das notícias que fizessem referência a um algoritmo baseado em conceitos pré-estudados, de modo a obter apenas os artigos relativos às minorias do domínio. Foi também desenvolvido um algoritmo de identificação e ordenação por camadas de prioridade em relação a cada minoria em estudo e peso das palavras-chave mencionadas, criando um sistema de pontuação do grau de proximidade e interesse de cada artigo em relação a cada domínio dentro das minorias, sistema esse implementado visualmente na plataforma do projeto.

Foram criadas interfaces e APIs para consulta pública dos dados. Todas as ferramentas estão disponibilizado em acesso aberto.

Objetivos

Os principais objetivos passam por disponibilizar uma ferramenta de investigação poderosa para a comunidade científica. O projeto compila e disponibiliza dados normalmente inacessíveis ao utilizador comum, devido à complexidade técnica envolvida nos processos informáticos relacionados com a mineração de dados, *Big Data* e ferramentas ontológicas de ponta. Para além da plataforma do projeto, mais acessível ao público geral, disponibilizamos *endpoints SPARQL* e APIs que permitem pesquisa e *download* facilitados. A título de exemplo, com uma simples *query* na plataforma do projeto, é possível obter uma tabela de dados com um conjunto de critérios complexos e extensos, de valor semântico profundo, por exemplo: "exibir todas as imagens que aparecem associadas a artigos publicados na primeira década do século XXI, que sejam relacionadas com as minorias 'mulheres' e mencionem a personalidade 'António Costa', mas a que tem um emprego como político não outro, e as palavras-chave 'proibição' e 'aborto', e que mencionem cidades de Portugal, excluindo Algarve (etc.)" - a *query* poderia ser estendida ao infinito. Numa base de dados tradicional este inter-relacionamento entre domínios semânticos, riqueza de resultados não seria possível com o mesmo grau de acessibilidade.

Este repositório de dados que pretende compilar e mapear a representação das minorias na imprensa portuguesa, foi especialmente idealizado para poder servir de material de estudo e pesquisa a diferentes grupos de investigação de diversas áreas, tais como Ciências Sociais, Humanidades, Estatística, Engenharia Informática, etc.

O *corpus* do projeto será atualizado e expandido anualmente com o material jornalístico do ano transato e novas fontes incrementadas (conforme disponibilização pelo Arquivo.pt).

Resultados Atingidos

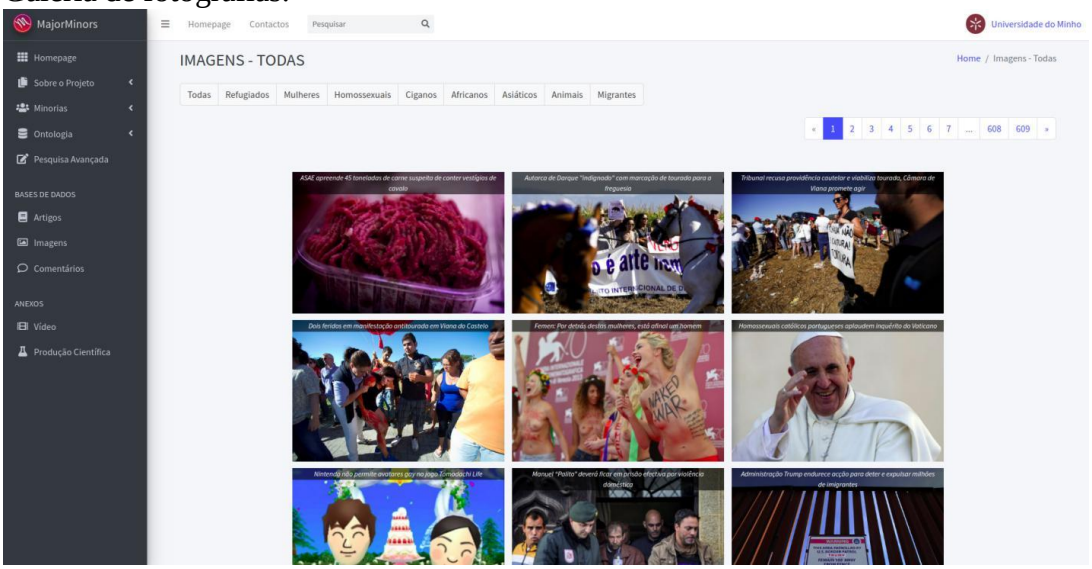
Neste momento a base de dados conta com 48.949 artigos, 3.348 comentários e 9.658 imagens, num total de 5.178.169 triplos ontológicos e referências cruzadas.

Destacamos o *website* do projeto, pois este pretende ser uma ferramenta de estudo no contexto das questões minoritárias:

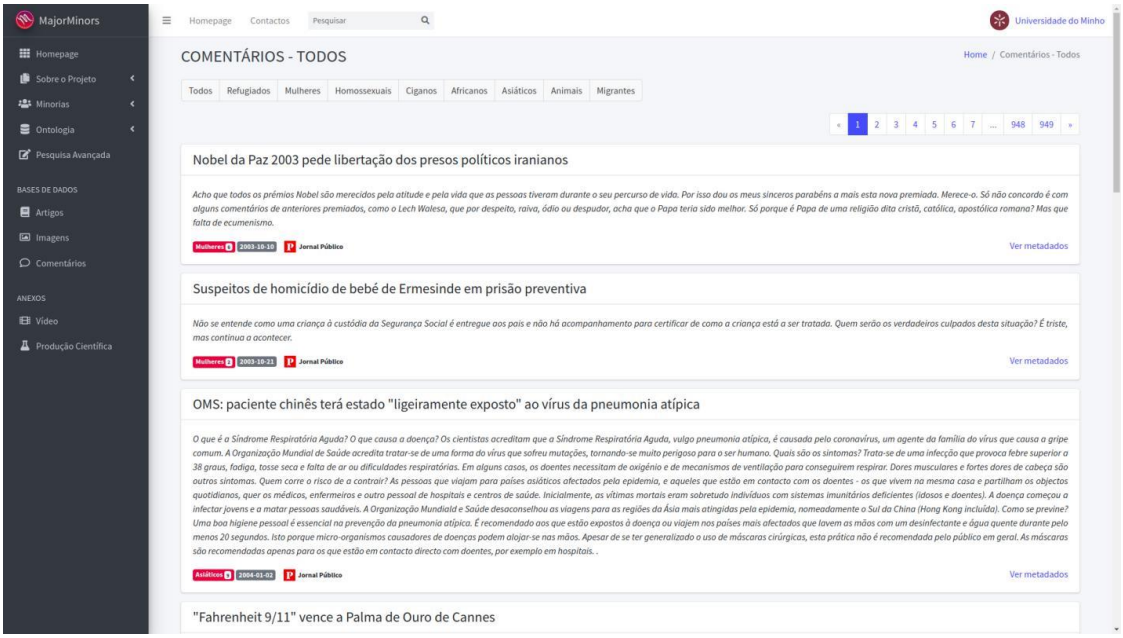
- http://minors.ilch.uminho.pt/

Este compila artigos, comentários e fotografias de recortes de imprensa que mencionam questões minoritárias (corpus: últimos 20 anos do jornal Público).

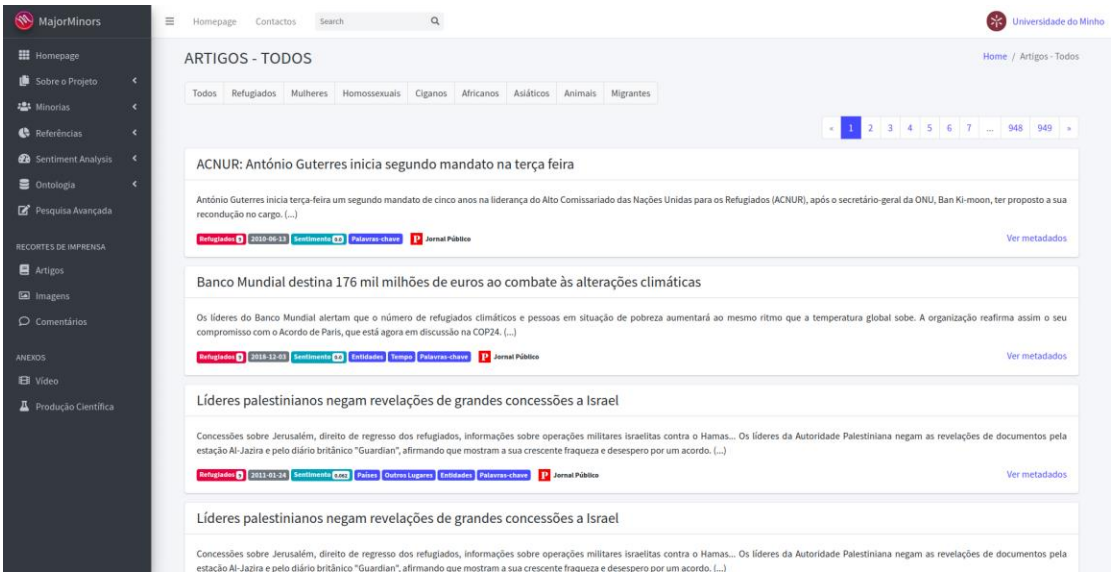
Galeria de fotografias:



Galeria de comentários de utilizadores:



Galeria de recortes de imprensa e referências cruzadas com outras entidades:



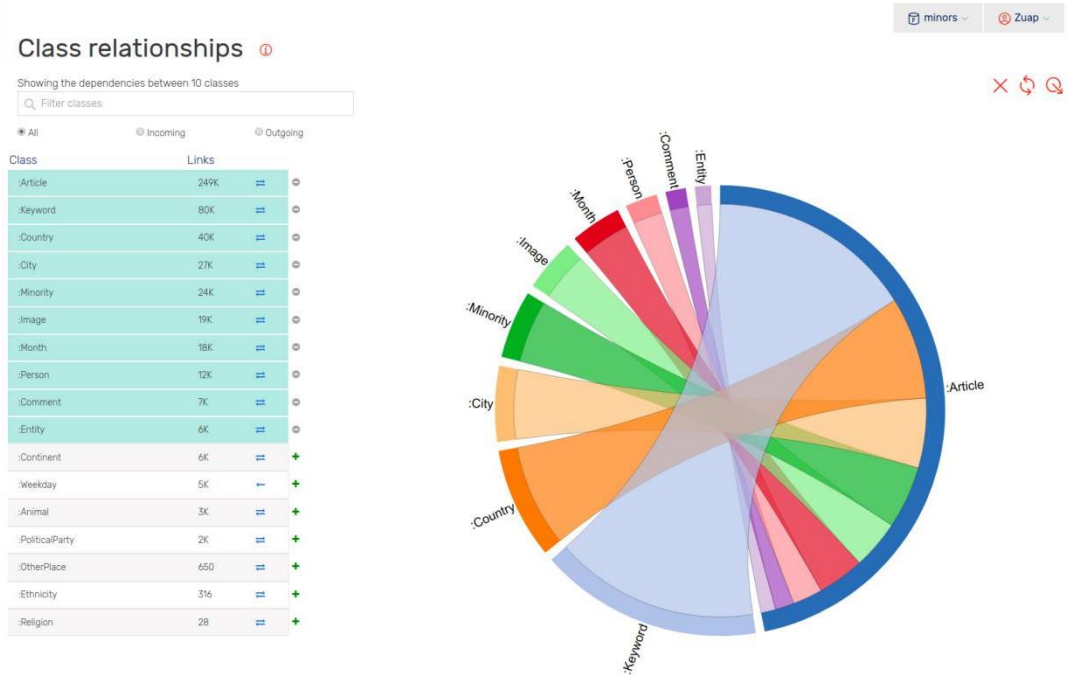
Estas galerias foram construídas para utilizadores sem conhecimentos técnicos de SPARQL e navegação em ontologias, mas são muito mais limitadas quanto aos dados que é possível cruzar. Numa segunda camada, construímos uma interface reativa (*VUE*) que permite navegar nos dados de forma dinâmica, oferendo dados em três níveis: classes ontológicas gerais, *individuals* no seu contra-domínimo, e filtros baseados em relações cruzadas básicas:

No entanto também esta camada é limitada comparando com as APIs SPARQL desenvolvidas. Essa API foi desenvolvida para um público-alvo técnico e académico, qualquer individuo interessado poderá facilmente manusear os dados de forma bastante complexa por meio de *queries* SPARQL. A documentação anexa é muito intuitiva. Esta especificação é uma recomendação da W3C para o futuro da *SemanticWeb*, sendo uma linguagem de *query* muito acessível mas ao mesmo tempo muito poderosa.

Destacamos um dos *endpoints SPARQL* configurados para essa navegação:

- http://sparql.ilch.uminho.pt

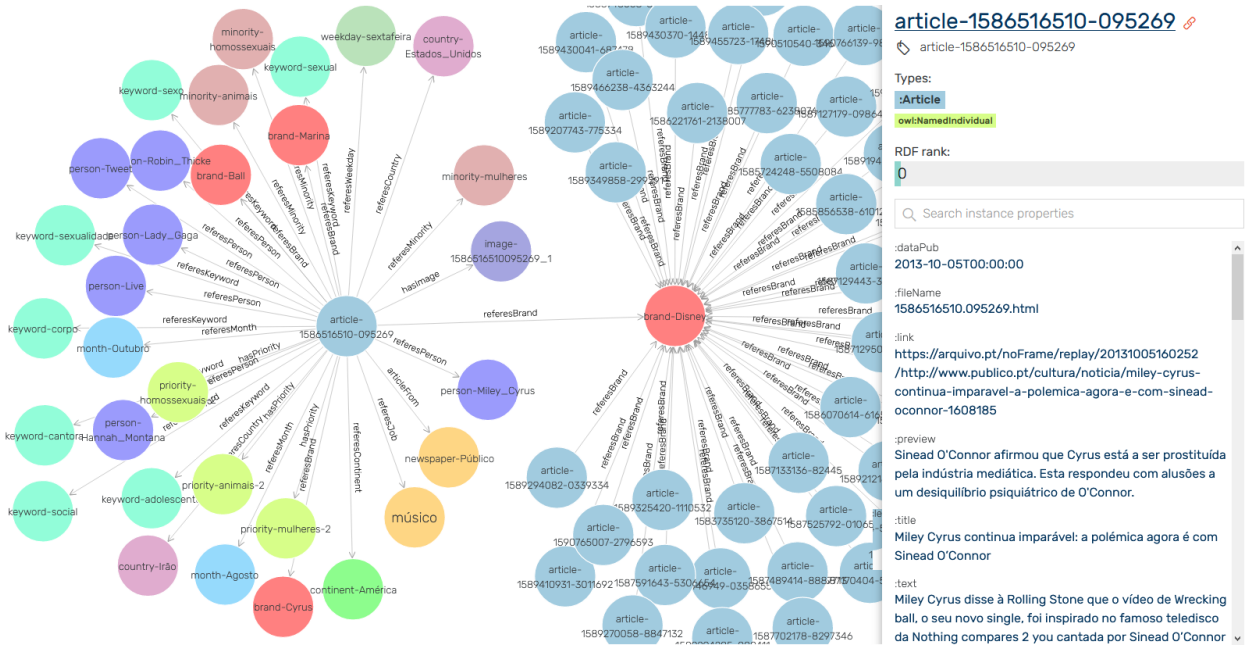
E a API e documentação em: http://minors.ilch.uminho.pt/sparql



Vivemos num contexto moderno em que acesso a *Big Data* é dispendioso e uma prioridade das empresas. Neste panorama, transformamos algo das elites técnicas num repositório gratuito em acesso aberto (licença de livre uso *MIT*). As ontologias utlizadas, nomeadamente *RDF*, *OWL* e *SPARQL*, estão catalogadas pelo *World Wide Web Consortium* como prioritárias para o desenvolvimento da *internet* e surgimento da *Web 3.0*, a *web* semântica.

Destacamos também alguns dados estatísticos no portal do projeto:

E interfaces de pesquisa SPARQL disponibilizadas, por exemplo:



Para além destes resultados tangíveis, o resultado mais importante não tangível é o **algoritmo** desenvolvido para identificação automática de textos com referência a questões minoritárias. Este mesmo **algoritmo**, para além de isolar as referências, também pontua a pertinência e intensidade da menção, podendo ser adaptado e facilmente utilizado noutros contextos (twittes, blogs, etc.). Estas ferramentas e documentação técnica encontram-se em acesso *open-source* nos repositórios mencionados nos recursos complementares.

Deixamos algumas tabelas que quantificam os resultados atingidos:

Table 1 Crawled output.

Type	Number
Newspapers articles (individual URIs)	94.309
Newspapers articles (non repeating titles)	48.949
Articles referring minorities (non repeating)	11.496
User comments related with minorities	3.348
Newspapers illustrations related with minorities	9.658

Table 2 Priority points associated with each article elements

Elements	Priority points
Main tile	7
Tags/Topics	5
Literal mention of the minority name	4
Subtitle/Headline	3
Preview	3
Corpus	2

Table 3 Entities extracted.

Type	No. of References
Public Figures	32.648
Political Parties	8.525
Minorities	11.530
Entities	11.473
Brands	46.629
Religions	260
Sports	4.333
Ethnicities	1.345
Car brands	969
Animals	6.199
TV Channels	3.799
Continents	7.119
Countries	28.794
Cities	28.933
Other Places	1.067
Newspapers defined tags	10.186
Weekdays	20.575
Months	24.113
Minorities related keywords	36.897

Table 4 Ontology output.

Type	Number
Ontology (All) triples	5.178.169
Infered (All) triples	2.084.435
Ontology (All) size	650,3 Mb
Ontology (Minorities) triples	1.110.411
Infered (Minorities) triples	453.330
Ontology (Minorities) size	125,7 Mb

Originalidade e caráter inovador

Este projeto é a primeira ontologia do género, nesta temática, em Portugal, pretendendo ser uma referência para investigação científica.

Uma ontologia é uma base de dados com representação semântica das relações entre conceitos. Podem ser usadas, entre outras coisas, com o objetivo de melhorar a exatidão de pesquisas, diminuindo o fosso homem-máquina. O *SPARQL* e *OWL* são linguagens de implementação desta tecnologia que tentam revolucionar a forma como interagimos com a Web. Por estas características, as ontologias são uma das tecnologias chaves para a implementação da nova geração da internet, a *Web* Semântica.

Impacto social

Para além da utilidade científica da mencionada anteriormente, a plataforma foi também idealizada para a utilização por parte de utilizadores comuns (não técnico-científicos). Para além da componente estatística e de base de dados, possui também uma vertente pedagógica - tem pois também o propósito de sensibilizar o leitor casual para as questões de igualdade e discriminação relacionadas com minorias. Para além das páginas técnicas, secções informativas no *website*, com o objetivo de introduzir as questões junto da população geral. A título de exemplo, tomamos a iniciativa de dar a conhecer a história por detrás de algumas das minorias, ou também algumas figuras relevantes pertencentes a cada uma delas que possivelmente o leitor não tinha conhecimento (como por exemplo, o facto de o Elvis Presley ter raízes na comunidade cigana):

Foi nosso objetivo produzir uma ferramenta útil à comunidade académica, mas também uma ferramenta pedagógica e de sensibilização extramuros à população em geral:

Impacto científico

Um dos motivos de avançar com o conceito do projeto foi o facto de vários grupos de investigação do Centro de Estudos Humanísticos da Universidade do Minho focarem temáticas próximas ao projeto, desde questões de Género, estudos sobre Identidade, análise de relações Homem e Animal, etc., no entanto, não há suporte técnico nem dados de fácil acesso que dinamizem estudos nestas áreas, havendo uma barreira técnica. Este projeto nasceu desse diálogo entre dois alunos de pós-graduação em Informática e as Ciências Humanas.

O projeto conta com a colaboração de Grupos de Investigação do Centro de Estudos Humanísticos da Universidade do Minho (CEHUM), nomeadamente o Grupo2i, colaborador principal, no âmbito de investigação de Estudos de Identidade, e o GHD, no âmbito do processamento e análise linguística, para além do apoio do Departamento de Informática da Universidade do Minho (DI). O objetivo do projeto passa por ser uma ferramenta de investigação para as áreas socioculturais e humanísticas, um *corpus* textual para análise linguística (principalmente no âmbito dos comentários dos utilizadores nas plataformas digitais dos jornais, que permitirá traçar perfis linguísticos e evolução da língua portuguesa) e, pela sua componente ontológica e *software* produzido para mineração/tratamento de dados, uma ferramenta informática.

A adição de dados complementares sobre notícias e elaboração de gráficos com base em dados estatísticos recolhidos foi idealizada com o propósito de fornecer acesso fácil a *Big Data* a utilizadores não técnicos.

Relevância da utilização do Arquivo.pt

O trabalho foi desenvolvido em redor de notícias arquivadas no *Arquivo.pt,* nomeadamente os últimos vinte anos de suporte jornalístico em formato digital jornal Público. Foi feito uso das *API's* disponíveis apara obter todas as páginas de entrada e subdomínios que foram recolhidos para alimentar o projeto, fazendo uso de *software* de *crawl* desenvolvido pela nossa equipa. O trabalho simplesmente não existiria sem o *Arquivo.pt*.

A i*nternet* é um novo conceito de biblioteca, contendo um imenso manancial de recursos, todos significativos para o perfil e retrato histórico da humanidade contemporânea. É, pois, de extrema urgência preservá-la, pois tal é sinónimo de preservar a nossa história. Com este projeto fomos capazes de criar um mapa da representação de minorias em contexto português nos últimos vinte anos e acompanhar a evolução e tendências. Sem o Arquivo.pt esta seria uma parte da nossa história perdida, uma oportunidade de investigação desperdiçada.

Recursos complementares

- **Website:** http://sparql.ilch.uminho.pt/sparql
- **Endpoint *SPARQL*:** http://sparql.ilch.uminho.pt/sparql
- **Galerias:**
  - **Artigos:** http://minors.ilch.uminho.pt/articles
  - **Fotos:** http://minors.ilch.uminho.pt/images
  - **Comentários:** http://minors.ilch.uminho.pt/comments
- **Interface reativa (pesquisa dinâmica):** http://minors.ilch.uminho.pt/search
- **API de consulta:** http://minors.ilch.uminho.pt/sparql
- **Estrutura da Ontologia (WebVowl):** http://minors.ilch.uminho.pt/ontology
- **Ferramentas Open-Source:**
  - **Crawler:** https://github.com/Paulo-Jorge-PM/crawler-majorminors
  - **Scrapper:** https://github.com/leandrocosta16/scraper-MajorMinors
  - **Entities Datasets:** https://github.com/Paulo-Jorge-PM/datasets-majorminors
  - **Ontology Assembler**: https://github.com/Paulo-Jorge-PM/ontology-assembler-majorminors
- **Relatório técnico:** https://www.overleaf.com/read/mvcngptpyckc