Development and exploration of an ontology for SAMH domain

Francisco Maciel
ei11084@fe.up.pt

Hugo Sousa
ei11083@fe.up.pt

Ricardo Silva
ei11079@fe.up.pt

January 8, 2016

Abstract

O SAMH é uma plataforma desenvolvida que permite perceber mais facilmente o estado emocional de um paciente, através da informação textual contida nas transcrições das suas consultas de psicologia. Seguindo os princípios da Semantic Web [1] e Open Linked Data [2] foi desenvolvida e explorada uma ontologia baseada no domínio desta plataforma, que se demonstrou ser benéfica e trazer valor adicional ao SAMH.

1 Introdução

O STOP DEPRESSION [3] é um projeto a decorrer atualmente em Portugal, que pretende promover uma melhoria da capacidade dos cuidados de saúde primários na prevenção, diagnóstico e tratamento da depressão e na prevenção do suicídio. Através deste projeto, obtiveram-se transcrições de consultas de psicologia relativas a 23 pacientes anónimos. O EMOTAIX.PT [4] é uma base lexical, traduzida de um projeto francês designado EMOTAIX [5], que fornece um mapeamento de termos para emoções, que se encontram divididas em 4 níveis hierárquicos.

Baseando-se nos *datasets* referidos, foi desenvolvida a plafatorma SAMH (*Sentiment Analysis for Mental Health*), que permite retornar as transcrições mais relevantes para um pesquisa por um conjunto de termos ou por um dado nível da base lexical. O *sentiment analysis* [6] contextualiza-se na origem desta plataforma, no sentido que através destas pesquisas torna-se mais fácil analisar e perceber o estado psicológico de um paciente.

Com base nos conceitos fundamentais da *Semantic Web*, foi desenvolvida uma ontologia baseada no domínio do problema, que poderá servir de suporte à plataforma SAMH. Assim sendo, a troca e reutilização de dados desta plataforma torna-se mais fácil e acessível, mesmo por outras máquinas.

2 Domínio do problema

Como se demonstra na figura 1 e tendo em conta os *datasets* disponíveis, é possível identificar entidades e relações entre estas no domínio do problema. As consultas de psicologia são um diálogo entre o terapeuta e o paciente, que são posteriormente transcritas num documento. Estas transcrições contêm múltiplos termos, que podem estar incluídos na base lexical. Um termo não é necessariamente precedido dos 4 níveis na base lexical, podendo suceder diretamente o nível primário.

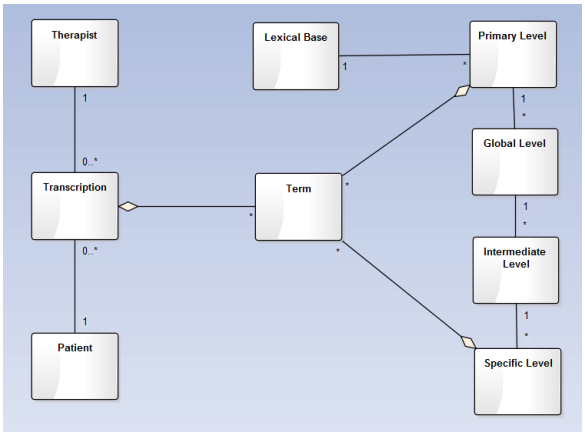


Figure 1: *Modelo conceptual do domínio do problema*

3 Ontologia

3.1 Análise e Pesquisa

O domínio do problema a ter em conta neste âmbito é consideravelmente reduzido e específico, pelo que se optou por desenvolver a ontologia de raiz. Uma breve análise de ontologias existentes não se demonstrou relevante para a especificidade deste domínio.

3.2 Ferramenta

Para o desenvolvimento da ontologia, foi usado o *protégé* [6], um editor de ontologias *open-source*.

3.3 Desenvolvimento

Apesar do reduzido domínio do problema, várias situações suscitaram discussão relativamente à forma de implementação da ontologia. Diferentes soluções discutem geralmente uma maior generalização ou restrição da ontologia ao modelo do problema. O diagrama final da ontologia está disponível no apêndice A.

3.3.1 Emotion Levels

Ao invés de existir uma entidade para representar cada nível emocional, poderia haver apenas uma entidade *LevelDegree*, com os respetivos 4 níveis, como se verifica na figura 2. Isto implicaria, no entanto, um conjunto de regras adicionais. Por exemplo, seria necessário indicar que apenas as instâncias *PRIMARY* e *SPECIFIC* podem ter termos. Também ao indicar que *EmotionLevel* é um nível superior de si próprio implicaria regras adicionais. Verifica-se aqui um *trade-off*. Por um lado, a solução final é mais restrita. No entanto, no caso de alguém querer reutilizar esta ontologia, as restrições impostas podem não corresponder ao modelo do problema em questão.

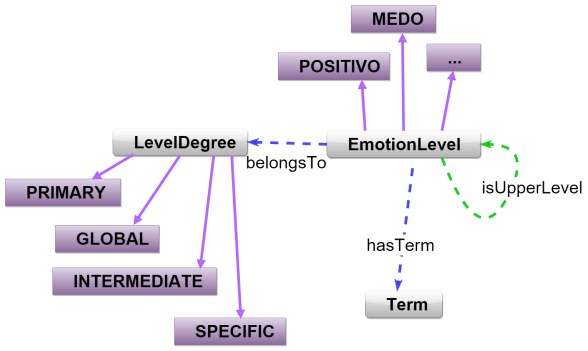


Figure 2: *Alternativa de representação de EmotionLevel*

3.3.2 Transcription Content

Verificando a figura 1, verificamos que se definiu uma transcrição como um conjunto de termos. A ideia inicial passaria por guardar todos os termos de uma transcrição, para posteriormente efetuar as relações com os termos da base lexical através de *queries* SPARQL. No entanto, esta é na verdade a funcionalidade que o *Solr* [7] já efetua, e a sua implementação através de *queries* seria complexa. Optou-se, então, por armazenar o conteúdo da transcrição como uma *data property* de *Transcription*, ao contrário do que se pode ver na alternativa apresentada figura 3. Esta alteração suscita a necessidade de representar a conexão entre o conteúdo de um documento e a base lexical.

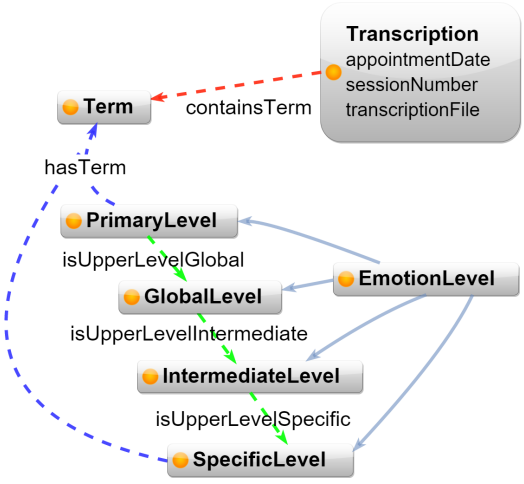


Figure 3: *Alternativa de representação do conteúdo de transcrição*

3.3.3 Emotional Conection

De forma a relacionar as transcrições a estados emocionais, seria uma opção armazenar apenas as *n* transcrições mais relevantes para um nível emocional. No entanto, haveria aqui muita informação perdida e que não seria possível ser explorada. A solução final guarda, para cada transcrição, o resultado para todos os níveis emocionais existentes. Esta solução é, obviamente, mais complexa espacialmente, mas potencia a ontologia com melhores resultados na sua exploração. Para representar o resultado de um nível emocional para uma transcrição, tendo em conta a limitação da *Semantic Web* que suporta apenas relações binárias [8], foi necessário criar uma nova entidade *SearchResult* , com *data property* "score", que é obtido através de uma *query* ao *Solr*. Ambas as representações dificultam a inserção de novos documentos, pois o "score" não é independente do universo de transcrições, carecendo todas as inserções da nova reavaliação por parte do *Solr* e repovoamento da ontologia.

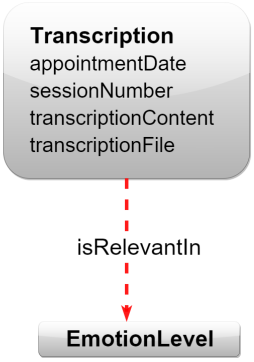


Figure 4: *Alternativa de representação da relação entre EmotionLevel e Transcription. Apenas um dado número de níveis emocionais relevantes para uma transcrição seriam armazenados.*

3.4 Povoamento

A ontologia foi povoada com todos os dados disponíveis. Para o povoamento ser realizado automaticamente, foram criados os seguintes módulos em Java e com recurso a *OWL API* [9], sendo que cada um deles é responsável pelo povoamento de diferentes dados:

- base lexical
- terapeutas, pacientes e transcrições
- conteúdo das transcrições
- *search results*

Os dois últimos módulos referidos requerem que o servidor *Solr* utilizado para o desenvolvido da plataforma SAMH esteja a correr localmente e respetivamente povoado. Com o *dataset* atual, este povoamento resulta num ficheiro final de 14.2Mb, em *OWL Functional Syntax* [10].

3.5 Exemplos de queries

Com recurso a SPARQL [11], é possível interrogar a ontologia desenvolvida. As seguintes *queries* são alguns exemplos de interrogações que poderiam ser interessantes no contexto da plataforma SAMH:

- pacientes atendidos por um dado terapeuta (ver apêndice B)
- níveis emocionais de um dado termo (ver apêndice C)
- termos de um dado nível emocional (ver apêndice D)
- melhor transcrição para um dado nível emocional (ver apêndice E)
- emoções de uma dada transcrição (ver apêndice F)
- emoções de um dado paciente (ver apêndice G)

As última duas *queries* referidas merecem uma nota adicional, pois o resultado obtido é algo que realmente traz valor acrescentado à plataforma. Uma pesquisa global das emoções é uma nova funcionalidade que tem potencial para ser bastante útil no contexto de *sentiment analysis*, possível devido à arquitetura escolhida na implementação da ontologia, mais concretamente na forma de armazenamento de *search results*.

4 Conclusões

Dado o modelo conceptual ser de dimensão reduzida, previa-se que o desenvolvimento da ontologia tivesse reduzida complexidade, o que não se sucedeu, suscitando a análise e discussão em variados tópicos. As funcionalidades e potencialidades da ontologia dependem fortemente do seu desenho inicial e, sem ter um conhecimento profundo do modelo conceptual, perceber as relações entre as diferentes entidades nem sempre é trivial.

A ferramenta utilizada demonstrou-se bastante funcional e relativamente fácil de usar, após entender os conceitos básicos das ontologias. A adaptação foi relativamente simples, pela sua relativa similaridade a bases de dados relacionais. Relativamente à visualização do grafo da ontologia após povoamento desta, torna-se impraticável devido ao elevado número de instâncias e propriedades.

A ontologia implementada mostrou-se ser útil no contexto da plataforma SAMH, trazendo valor adicional a esta e contribuindo para uma positiva implementação de um caso de uso na *Semantic Web*.

A interrogação da ontologia mostrou-se competente e exequível numa plataforma real. A ontologia poderia substituir uma eventual base de dados. O principal problema das ontologias é a sua escalabilidade, mas soluções têm vindo a ser apresentadas e discutidas [12, 13].

O domínio do problema apresentado é bastante dinâmico. Assim sendo, seria necessária uma atualização constante da ontologia. Por exemplo, ao inserir uma transcrição, todos os *search results* devem ser atualizados, pois o *score* resultante da *query* ao *Solr* depende também do *document frequency*. Esta atualização deverá ser demorada dado o elevado número de *object properties* a modificar, poderia ser efetuada diariamente ou por pedido de um utilizador com permissões para tal.

References

[1] W3C. W3C Semantic Web Activity. http://www.w3.org/2001/sw/. Acedido em Janeiro, 2015.

[2] W3C. Linked Data. http://www.w3.org/standards/semanticweb/data. Acedido em Janeiro, 2015.

[3] STOP DEPRESSION. https://stopdepression.ismai.pt/. Acedido em Janeiro, 2015.

[4] Sara Costa, Rui A. Alves, Fernando Barbosa, e Thierry Olive. EMOTAIX.PT, an emotional word database in European Portuguese. Em *SIG Writing Porto 2012*, número 13 em International Conference of the EARLI Special Interest Group on Writing. Universidade do Porto, Julho 2012.

[5] Annie Piolat e Rachid Bannour. An example of text analysis software (emotaix-tropes) use: The influence of anxiety on expressive writing. *Current psychology letters, Vol. 25, Issue 2*, páginas 91–108, 2003.

[6] Stanford Center for Biomedical Informatics Research. protégé. http://protege.stanford.edu/. Acedido em Janeiro, 2015.

[7] The Apache Software Foundation. Apache Solr. http://lucene.apache.org/solr/. Acedido em Janeiro, 2015.

[8] W3C. Defining N-ary Relations on the Semantic Web. http://www.w3.org/TR/swbp-n-aryRelations/. Acedido em Janeiro, 2015.

[9] The OWL API. http://owlapi.sourceforge.net/. Acedido em Janeiro, 2015.

[10] W3C. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition). http://www.w3.org/TR/owl2-syntax/. Acedido em Janeiro, 2015.

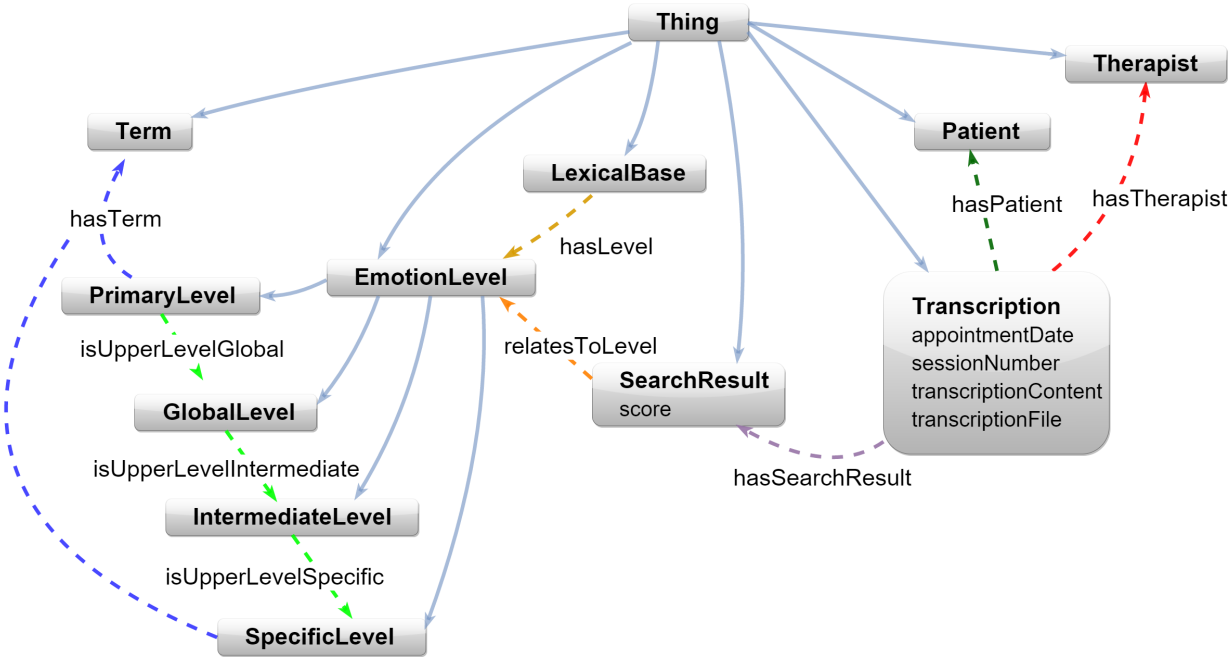
[11] W3C. SPARQL 1.1 Overview. http://www.w3.org/TR/sparql11-overview/. Acedido em Janeiro, 2015.

[12] Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani Thuraisingham, e Paolo Castagna. Jena-HBase: A Distributed, Scalable and Efficient RDF Triple Store. Relatório técnico.

[13] Jiewen Huang, Daniel J. Abadi, e Kun Ren. Scalable SPARQL Querying of Large RDF Graphs. *PVLDB*, 4(11):1123–1134, 2011.

Appendices

A Grafo da ontologia



B Query - Pacientes atendidos por um dado terapeuta

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns\#>
PREFIX owl: <http://www.w3.org/2002/07/owl\#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema\#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema\#>
PREFIX samh: <http://www.semanticweb.org/SAMH/ontologies/>

SELECT DISTINCT (?p as ?Patient)
WHERE {
        ?t a samh:Transcription .
        ?t samh:hasPatient ?p .
        ?t samh:hasTherapist ?therapist
        filter( regex(str(?therapist), "Salgado"))
}
ORDER BY ?Patient
```

C Query - Níveis emocionais de um dado termo

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX samh: <http://www.semanticweb.org/SAMH/ontologies/>

SELECT DISTINCT ?pLevel ?gLevel ?intLevel ?spLevel
WHERE {
        ?x a samh:Term .
        {
                ?spLevel samh:hasTerm ?x .
                ?intLevel samh:isUpperLevelSpecific ?spLevel .
                ?gLevel samh:isUpperLevelIntermediate ?intLevel .
                ?pLevel samh:isUpperLevelGlobal ?gLevel .
        }
        UNION
         {
                ?pLevel samh:hasTerm ?x .
                FILTER NOT EXISTS  { ?upperLevel samh:isUpperLevelSpecific ?pLevel
        }
        filter( regex(str(?x), "neutro"))
}
```

D Query - Termos de um dado nível emocional

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX samh: <http://www.semanticweb.org/SAMH/ontologies/>

SELECT DISTINCT (?t as ?terms)
WHERE {
        {
                ?eLevel samh:hasTerm ?t
        }
        UNION
        {
                ?eLevel samh:isUpperLevelSpecific ?spLevel .
                ?spLevel samh:hasTerm ?t
        }
        UNION
        {
                ?eLevel samh:isUpperLevelIntermediate ?intLevel .
                ?intLevel samh:isUpperLevelSpecific ?spLevel .
                ?spLevel samh:hasTerm ?t
        }
        UNION
        {
                ?eLevel samh:isUpperLevelGlobal ?gLevel .
                ?gLevel samh:isUpperLevelIntermediate ?intLevel .
```

```
                ?intLevel samh:isUpperLevelSpecific ?spLevel .
                ?spLevel samh:hasTerm ?t
        }
        filter( regex(str(?eLevel), "AMOR" ))
}
ORDER BY ?t
```

E Query - Melhor transcrição para um dado nível emocional

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX samh: <http://www.semanticweb.org/SAMH/ontologies/>

SELECT DISTINCT (?t as ?Transcription) ?score
WHERE
{
        ?sr rdf:type samh:SearchResult .
        ?sr samh:relatesToLevel ?l .
        ?t samh:hasSearchResult ?sr .
        ?sr samh:score ?score .
        filter( regex(str(?l), "AMOR" ))
}
ORDER BY DESC(?score)
LIMIT 1
```

F Query - Emoções de uma dada transcrição

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX samh: <http://www.semanticweb.org/SAMH/ontologies/>

SELECT DISTINCT (?l as ?EmotionLevel) ?score
WHERE
{
        ?sr rdf:type samh:SearchResult .
        ?sr samh:relatesToLevel ?l .
        ?t samh:hasSearchResult ?sr .
        ?sr samh:score ?score .
        filter( regex(str(?t), "P003_13" ))
}
ORDER BY DESC(?score)
```

G Query - Emoções de um dado paciente

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX samh: <http://www.semanticweb.org/SAMH/ontologies/>

SELECT DISTINCT (?l as ?EmotionLevel) (SUM(?score) as ?sum)
WHERE
{
        {
                SELECT DISTINCT ?t ?p
                WHERE
                {
                        ?t a samh:Transcription .
                        ?t samh:hasPatient ?p .
                        filter( regex(str(?p), "P003" ))
                }
        }

        ?sr rdf:type samh:SearchResult .
        ?sr samh:relatesToLevel ?l .
        ?t samh:hasSearchResult ?sr .
        ?sr samh:score ?score .
}
GROUP BY ?l
ORDER BY DESC(?sum)
```