UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

Semantic Annotation of Clinical Questionnaires to Support Personalized Medicine

André Gonçalves

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof. Doutora Cátia Luísa Santana Calisto Pesquita
Prof. Doutor Tiago João Vieira Guerreiro

2021

Acknowledgements

First of all, I would like to acknowledge my supervisor Professor Cátia Pesquita who was relentless in the support given and her level of excellence and exigency allowed me to be one level above what I was used to. A word of thanks also to my co-advisor Professor Tiago Guerreiro for the availability and support demonstrated.

A big thank you to all doctors and therapists of the Senior Neurological Campus who actively participated in this dissertation and were allowed to confer a scientific value to the Thesis that without them would not be equal, thank you CNS Clinical Director, Dr Joaquim Ferreira, Dr Linda Kauppila, nurse Marta Pires, physiotherapist Filipa Pona, therapists Rafaela Dias, Mariana Morgado and Joana Ramalho, nutritionist Joana Malheiro and psychologist Maria João Silvestre.

I would like to dedicate this dissertation to my parents and sister who supported and accompanied me from the beginning of everything and watched my evolution to which they contributed a lot with their patience and understanding, especially throughout this journey in which they had to endure my crazy study schedule.

To my friends Ricardo and Susana for sharing with me all the moments of this academic journey and for all the incredible moments that we spent inside and outside the master's degree.

Thank you to all my friends that FCUL gave me and have accompanied me since graduation until today, especially Afonso for all the hours of discussion about the most diverse topics, my Hermarcita who appeared late but with a huge and lovely impact, and Leonor my partner of a countless number of adventures and hours of conversation.

To my friends of a lifetime from Odivelas, who have been by my side since I can remember and on whom I could always count, Duarte, Miguel, Bernardo, Rodrigues and Tomás a little of this thesis also belongs to you for the unconditional support with which you never failed me.

To my family...

Resumo

Atualmente estamos numa era global de constante evolução tecnológica, e uma das áreas que têm beneficiado com isso é a medicina, uma vez que com integração da vertente tecnológica na medicina, tem vindo a ter um papel cada vez mais importante quer do ponto de vista dos médicos quer do ponto de vista dos pacientes.

Como resultado de melhores ferramentas que permitam melhorar o exercício das funções dos médicos, estão se a criar condições para que os pacientes possam ter um melhor acompanhamento, entendimento e atualização em tempo real da sua condição clínica.

O setor dos Cuidados de Saúde é responsável pelas novidades que surgem quase diariamente e que permitem melhorar a experiência do paciente e o modo como os médicos podem tirar proveito da informação que os dados contêm em prol de uma validação mais célere e eficaz. Este setor tem gerado um volume cada vez mais maciço de dados, entre os quais relatórios médicos, registos de sensores inerciais, gravações de consultas, imagens, vídeos e avaliações médicas nas quais se inserem os questionários e as escalas clínicas que prometem aos pacientes um melhor acompanhamento do seu estado de saúde, no entanto o seu enorme volume, distribuição e a grande heterogeneidade dificulta o processamento e análise.

A integração deste tipo de dados é um desafio, uma vez que têm origens em diversas fontes e uma heterogeneidade semântica bastante significativa; a integração semântica de dados biomédicos resulta num desenvolvimento de uma rede semântica biomédica que relaciona conceitos entre diversas fontes o que facilita a tradução de descobertas científicas ajudando na elaboração de análises e conclusões mais complexas para isso é crucial que se atinja a interoperabilidade semântica dos dados. Este é um passo muito importante que permite a interação entre diferentes conjuntos de dados clínicos dentro do mesmo sistema de informação ou entre sistemas diferentes. Esta integração permite às ferramentas de análise e interface com os dados trabalhar sobre uma visão integrada e holística dos dados, o que em última análise permite aos clínicos um acompanhamento mais detalhado e personalizado dos seus pacientes.

Esta dissertação foi desenvolvida no LASIGE e em colaboração com o Campus Neurológico Sénior e faz parte de um grande projeto que explora o fornecimento de mais

e melhores dados tanto a clínicos como a pacientes. A base deste projeto assenta numa aplicação web, o DataPark que possui uma plataforma que permite ao utilizador navegar por áreas clinicas entre as quais a nutrição, fisioterapia, terapia ocupacional, terapia da fala e neuropsicologia, em que cada uma delas que alberga baterias de testes com diversos questionários e escalas clínicas de avaliação. Este tipo de avaliação clínica facilita imenso o trabalho do médico uma vez que permite que sejam implementadas à distância uma vez que o paciente pode responder remotamente, estas respostas ficam guardadas no DataPark permitindo ao médico fazer um rastreamento do status do paciente ao longo do tempo em relação a uma determinada escala.

No entanto o modo como o DataPark foi desenvolvido limita uma visão do médico orientada ao questionário, ou seja o médico que acompanha o paciente quando quer ter a visão do mesmo como um todo tem esta informação espalhada e dividida por estes diferentes questionários e tem de os ir ver a todos um a um para ter a noção do status do paciente. Esta dissertação pretende fazer face a este desafio construindo um algoritmo que decomponha todas as perguntas dos diferentes questionários e permita a sua integração semântica. Isto com o objectivo de permitir ao médico ter um visão holística orientada por conceito clínico.

Procedeu-se então à extração de toda a base de dados presente no DataPark, sendo esta a fonte de dados sobre a qual este trabalho se baseou, frisando que originalmente existem muitos dados em Português que terão de ser traduzidos automaticamente.

Com uma análise de alto nível (numa fase inicial) sobre os questionários da base de dados, iniciou-se a construção de um modelo semântico que pudesse descrever os dados presentes nos questionários e escalas. Assim de uma forma manual foi feito um levantamento de todos os conceitos clínicos que se conseguiu identificar num sub-conjunto de questionários, mais concretamente 15 com os 5 mais respondidos em relação à Doença de parkinson, os 5 mais respondidos em relação à doença de AVC e os 5 mais respondidos que não estejam associados a uma única patologia em específico. Este modelo foi melhorado e evoluiu em conjunto com uma equipa de 12 médicos e terapeutas do CNS ao longo de 7 reuniões durante as quais foi levado a cabo um workshop de validação que permitiu dotar o modelo construído de uma fiabilidade elevada.

Em paralelo procedeu-se à elaboração de 2 estudo: (i) um estudo que consistia em avaliar com qual ou quais ontologias se obtém a maior cobertura dos dados do sub-conjunto de 15 questionários. A conclusão a que se chegou foi que o conjunto de ontologias que nos conferia mais segurança é constituído pelas ontologias LOINC, NCIT, SNOMED e OCHV, conjunto esse foi utilizado daqui em diante; (ii) outro estudo procurou aferir qual a ferramenta de tradução automática(Google Translator ou Microsoft Translator) que confere uma segurança maior, para isso procedeu-se à tradução completa de 3 questionários que apesar de estar na base de dados no idioma português, tem a sua versão original em inglês. Isto permitiu-nos traduzir estes 3 questionários de português para inglês e avaliar em qual das duas ferramentas se obteve uma melhor performance. O Microsoft Translator apresentou com uma diferença pequena um desempenho superior, sendo portanto a ferramenta de tradução automática escolhida para integrar o nosso algoritmo.

Concluídos estes 2 estudos temos assim o conjunto de dados uniformizado numa só linguagem, e o conjunto de ontologias escolhidas para a anotação semântica. Para entender esta fase do trabalho há que entender que ontologias são poderosas ferramentas computacionais que consistem num conjunto de conceitos ou termos, que nomeiam e definem as entidades presentes num certo domínio de interesse, no ramo da biomedicina são designadas por ontologias biomédicas.

O uso de ontologias biomédicas confere uma grande utilidade na partilha, recuperação e na extração de informação na biomedicina tendo um papel crucial para a interoperabilidade semântica que é exatamente o nosso objectivo final.

Assim sendo procedeu-se à anotação semântica das questões do sub-conjunto de 15 questionários, uma anotação semântica é um processo que associa formalmente o alvo textual a um conceito/termo, podendo estabelecer desta forma pontes entre documentos/texto-alvos diferentes que abordam o mesmo conceito. Ou seja, uma anotação semântica é associar um termo de uma determinada ontologia a um conceito presente no texto alvo. Imaginando que o texto alvo são diferentes perguntas de vários questionários, é natural encontrar diferentes questões de diferentes áreas de diagnóstico que estejam conectados por termos ontológicos em comum.

Depois da anotação completada é feita a integração do modelo semântico, com o algoritmo desenvolvido com o conjunto de ontologias e ainda com os dados dos pacientes. Desta forma sabemos que um determinado paciente respondeu a várias perguntas que abordam um mesmo conceito, essas perguntas estão interligadas semanticamente uma vez que têm o mesmo conceito mapeado.

A nível de performance geral tanto os processos tradução como de anotação tiveram um desempenho aceitável, onde a nivel de tradução se atingiu 78% *accuracy*, 76% *recall* e uma *F-mesure* de 0.77 e ao nível da performance de anotação obteve-se 87% de anotações bem conseguidas. Portanto num cômputo geral consegue-se atingir o principal objectivo que era a obtenção holística integrada com o modelo semântico e os dados do DataPark(Questionários e pacientes).

Palavras-chave: Anotação Semântica, Integração Semântica, Modelo Semântico, Questionários clínicos, Escalas Clínicas, Tradução automática.

Abstract

Healthcare is a multi-domain area, with professionals from different areas often collaborating to provide patients with the best possible care. Neurological and neurodegenerative diseases are especially so, with multiple areas, including neurology, psychology, nursing, physical therapy, speech therapy and others coming together to support these patients.

The DataPark application allows healthcare providers to store, manage and analyse information about patients with neurological disorders from different perspectives including evaluation scales and questionnaires. However, the application does not provide a holistic view of the patient status because it is split across different domains and clinical scales.

This work proposes a methodology for the semantic integration of this data. It developed the data scaffolding to afford a holistic view of the patient status that is concept-oriented rather than scale or test battery oriented. A semantic model was developed in collaboration with healthcare providers from different areas, which was subsequently aligned with existing biomedical ontologies. The questionnaire and scale data was semantically annotated to this semantic model, with a translation step when the original data was in Portuguese. The process was applied to a subset of 15 scales with a manual evaluation of each process. The semantic model includes 204 concepts and 436 links to external ontologies. Translation achieved an accuracy of 78%, whereas the semantic annotation achieved 87%. The final integrated dataset covers 443 patients.

Finally, applying the process of semantic annotation to the whole dataset, conditions are created for the process of semantic integration to occur, this process consists in crossing all questions from different questionnaires and establishing a connection between those that contain the same annotation.

This work allows healthcare providers to assess patients in a more global fashion, integrating data collected from different scales and test batteries that evaluate the same or similar parameters.

Keywords: Semantic Annotation, Semantic Integration, Semantic Model, Clinical Questionnaires, Clinical Scales, Machine Translation.

Contents

Figures Contents

Table Contents

Chapter 1

Introduction

In recent years technologies have been increasingly integrated into medicine, playing a more and more important role both from the point of view of doctors and patients. By defining better tools from which doctors can take advantage, making the exercise of their functions more dynamic, conditions are also created so that patients can have a better follow-up, understanding and updating of their clinical condition (Sreeninvasan and Chacko, 2020). All this growth around the evolution of clinical assessment methods contributes to sustaining the increasingly rich healthcare domain.

The healthcare sector in the last decade has undergone a large and accelerated growth and such evolution translates into a massive growth in the volume of clinical data produced (Dhayne *et al.*, 2019; le Sueur *et al.*, 2020), among which medical reports, inertial sensor records, appointment recordings, images, videos and medical assessments in which questionnaires and clinical scales are included (Dugas *et al.*, 2016; Sreeninvasan and Chacko, 2020). The consequence of this huge volume of data is a great heterogeneity that hinders its process and rapid analysis (Dhayne *et al.*, 2019).

1.1 Motivation and Context

Healthcare is a multi-domain area, with professionals from different areas often collaborating to provide patients with the best possible care. Neurological and neurodegenerative diseases are especially so, with multiple areas, including neurology, psychology, nursing, physical therapy, speech therapy and others coming together to support these patients.

This master thesis, which was developed at LASIGE[1] and in collaboration with the Campus Neurológico Senior (CNS)[2], is part of a project exploring the provision of more and better data to both clinicians and patients in the area of neurological degenerative

[1] https://www.lasige.pt/

[2] https://www.cnscampus.com/en

diseases. This project builds DataPark, a web application that allows clinicians to obtain more information about patients from different perspectives such as assessment through objective data (free-living) and subjective data. Objective data is obtained through inertial sensors while subjective data is obtained by filling in electronic records, where the standardized clinical questionnaires also referred to as clinical scales are inserted (Branco et al., 2019). DataPark also records the results of standardized clinical questionnaires and clinical scales, essential tools in the evaluation of various clinical parameters and work based on scores given by the answers given by the patient where the final score allows the physician to have an idea of the patient's status about the object of study of the scale/questionnaire. These scales provide quality and confidence since they have a research base behind their development (Zapata-Ospina and García-Valencia, 2020). This data is the focus of this dissertation.

The organisation of these questionnaires in DataPark is done in 5 different main areas, nutrition, physiotherapy, occupational therapy, speech therapy and neuropsychology, all independently assessed diagnostic areas (Figure 1). Each of these areas has several batteries of tests where the questionnaires are then inserted.

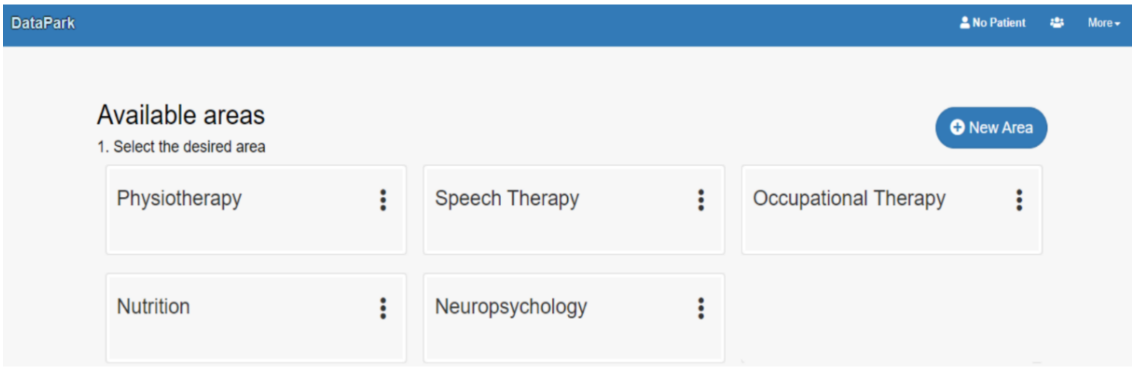
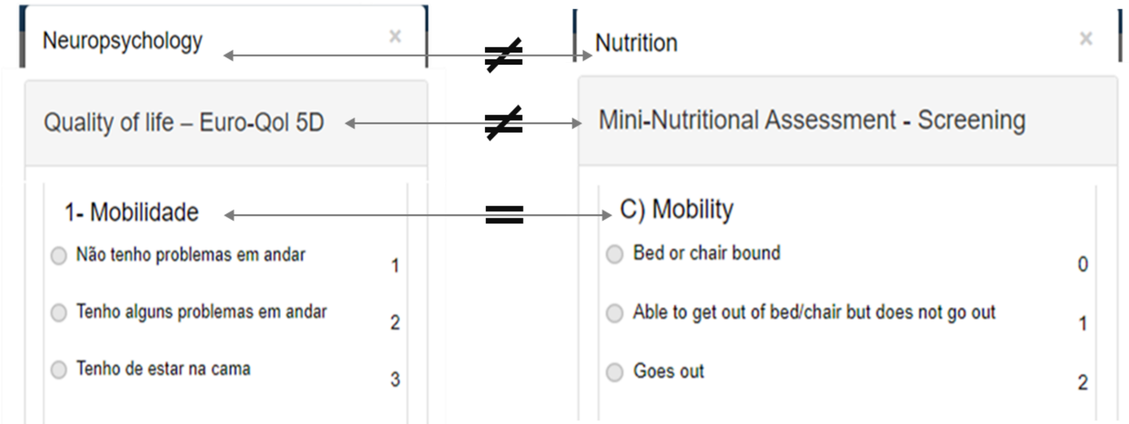


Figure 1: Window of the DataPark platform showing the 5 main diagnostic areas.

DataPark also allows the clinician to add new scales or questionnaires that the clinician considers pertinent to support the assessment of the patient about a specific pathology. These questionnaires allow the doctors to get an insight of the patient out of a controlled environment. The patient can answer these questionnaires either in the presence of the doctor or outside the doctor's office since DataPark allows answering the questionnaires through a web application or by phone way (IVR system). This way, the physician can schedule the execution of a given questionnaire throughout the day, week or month and evaluate the answers on the same scale over time.

However, the way DataPark was built limits the evaluation of the doctor-oriented questionnaire, and since the questionnaires are prepared and/or applied by different specialists, the doctor who follows the patient when he wants to have a view of the patient as a whole has this information scattered and divided by these different questionnaires and has to go see them all one by one to have a global notion of the patient's status. This is the main point of this dissertation if an algorithm can be built that deconstructs all the questions from different questionnaires and allows the semantic integration of them? This would allow the doctor to have a vision guided by the clinical concept and thus consult all the questions that assess a certain status, symptom, pathology, etc.

Figure 2 illustrates the current way of evaluating the content present in the questionnaires, in this specific example there are two questions present in the DataPark clinical questionnaires, with origin in different test batteries and diagnostic areas (Nutrition and Neuropsychology), however, they address the same subject, mobility. In cases like the example, it is useful to perform semantic integration. This example shows another challenge, the existence of questionnaires in different languages.



Neuropsychology ×

Quality of life – Euro-Qol 5D

1- Mobilidade

○ Não tenho problemas em andar 1

○ Tenho alguns problemas em andar 2

○ Tenho de estar na cama 3

Nutrition ×

Mini-Nutritional Assessment - Screening

C) Mobility

○ Bed or chair bound 0

○ Able to get out of bed/chair but does not go out 1

○ Goes out 2

Figure 2: Example of the appearance of the questions present in the data. Two questions address the same concept.

In short, DataPark questionnaires are transversal to several areas of medicine, their content being dispersed over several domains of biomedicine, and the different languages add to the difficulty in integrating these clinical data and giving healthcare providers a holistic view of the patient status

1.2 Objectives

The goal of this dissertation is to integrate data about a patient across different domains and clinical scales evaluation to afford a holistic view of the patient status that is concept-oriented rather than scale or test battery oriented. This will allow healthcare providers to assess patients in a more global fashion, integrating data collected from different scales and test batteries that evaluate the same or similar parameters.

To fulfil this goal, this work proposes to establish semantic annotations of the clinical scales and questionnaires that support the management, integration and analysis of the heterogeneous data available on Datapark to support the future development of new applications and interfaces for the assessment and monitoring of the clinical evolution of neurodegenerative and neurological patients in an integrated way.

The integration of the clinical questionnaires using ontology-based annotation allows for an (i) formal description of their addressed concepts/terms, even if coming from different medical and therapeutic specialities, and the (ii) association of different questionnaires to the same concept/term, through ontologies that establish an association relation among them.

1.3 Research questions

The work is organised to answer 3 main research questions that I intend to see answered throughout the dissertation:

- Q-1: How can a semantic model be established to describe the clinical information encoded in assessment questionnaires?

- Q-2: Can semi-automatic methods be used to integrate data from questionnaires using the semantic model?

- Q-3: What level of support does semantically annotated data provide to clinicians in their assessment?

All three issues are related but each has specific challenges. The construction of the semantic model involves prior data harmonization, which is a time-consuming step since it always requires manual evaluation and interaction with experienced experts in the field.

The choice of the right ontologies to integrate into the model is central for the correct description of the associated clinical information (Q-1); the choice of the automatic methods to be used is critical for the successful integration of the questionnaires from the model (Q-2): finally, obtaining an integrated view of the different associated questionnaires allows physicians a faster evaluation, which is only possible if the semantic integration of the clinical data is successful (Q-3).

1.4 Contributions

The main contributions are described into 3 below points:

- T model that describes the data present in the DataPark questionnaires, and which is faithful to the mental model of the CNS doctors, an involving and crucial part in the development of this semantic model.
- Methodology semantic that allows from a Database and a semantic model to automatically generate semantically linked data.
- The algorithm that allows the integration of different questionnaires with the supplied patient data, generating a holistic view of the patient-oriented to a certain medical concept.

1.5 Document structure

Starting with this chapter, which provides a contextualization of the problem faced and the proposed solution, this document has five more chapters structured as follows:

- **Chapter 2 (Concepts):** Explains the basic concepts for the understanding of this dissertation.
- **Chapter 3 (Related work)**: Exposure of recent works that are identified with the work done in this dissertation.
- **Chapter 4 (Design and Implementation)**: General presentation of the methodology followed and the algorithm developed.
- **Chapter 5 (Results and Discussion)**: Results obtained through the methodology developed in chapter 4.
- **Chapter 6 (Conclusion)**: Main conclusions from this work and proposals for future work.

Chapter 2

Concepts

In this section, I will set out the core concepts to understand the study done with this dissertation.

2.1 Ontologies

An ontology can be simply defined as an explicit specification of a conceptualization (Gruber, 1995). An ontology also defines the relationship between concepts in web documents. It enables machines to understand and process relevant documents and facilitate information sharing (Omid Yousefianzadeh, 2020). On a computational level, ontologies provide a amenable description of the set of concepts in a domain of interest and how they relate to each other (Hoehndorf, Schofield and Gkoutos, 2015), allowing to develop algorithms and decision systems that take advantage of them.

An ontology with a good level of comprehensiveness should have classes and relationships that model a given domain representing a shared detailed set of knowledge of that domain. In this field, the base of some reasoning about a specific domain is mainly formed by concepts (or classes) and relationships (Figure 3). That being said, in biomedical informatics, ontologies play a crucial role, since they allow interoperability across several systems, and support the integration of heterogeneous data sources.

Ontologies in biomedicine, called biomedical ontologies, are increasingly common. Currently, there are more than 900 ontologies available in BioPortal[3], a repository of biomedical ontologies (Amith *et al.*, 2018). Biomedical ontologies have an increasing number of fields of action in biomedicine, making it possible, for example, the integration of data that previously, despite being related, would have been captured independently and unrelated to each other. Currently, they play a strong role in the biomedical context(Omid Yousefianzadeh, 2020), where the volume of data that is produced is massive and with immense possibilities for bridging data (Hoehndorf, Schofield and Gkoutos, 2015) including:

- Aspire to a much more complete knowledge network, or when integrated into electronic health records provide the possibility of new methods of classification and stratification of patients;

[3] https://bioportal.bioontology.org/

- The analysis and mining of large-scale patient data;
- The use of ontology-based enrichment algorithms on data such as exomes and whole-genome sequences;
- New methods for incorporating results from biological research, enhancing improvements in clinical decision-making;
- Collect knowledge-rich information that is related to a common point, e.g., medical procedures, drugs, diseases, and genotypes are independent but strongly connected fields of knowledge, a connection that biomedical ontologies can easily make.

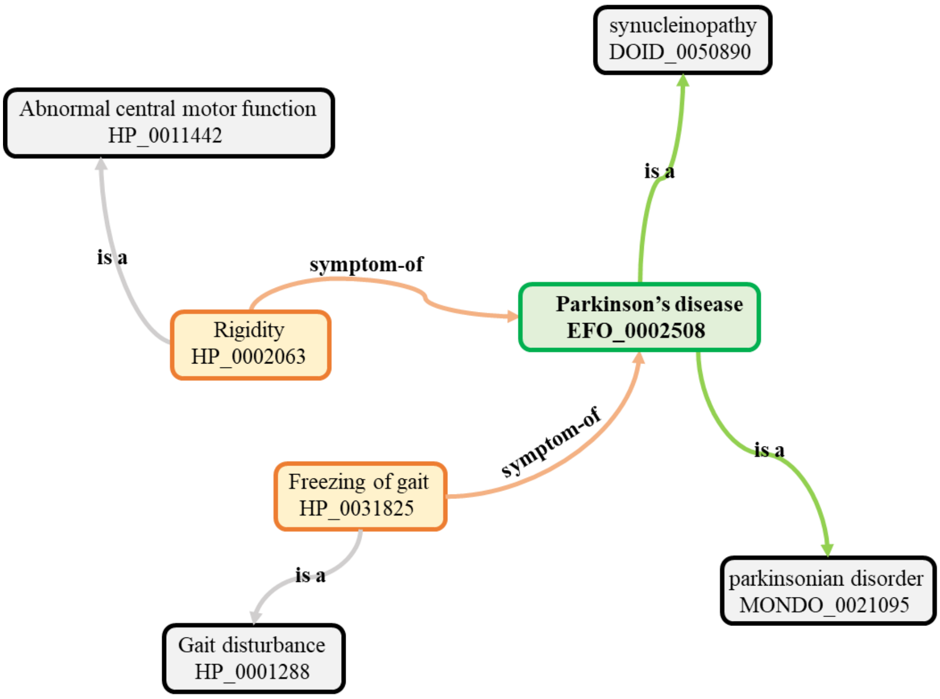


Figure 3: Example of a relationship between different entities coming from different ontologies such as EFO, disease ontology and human phenotype ontology that establish relationships between Parkinson's disease, its symptoms, and its classifications.

Biomedical ontologies were developed with the main goal to face the great demand for categorizing, reusing and sharing biomedical data. An unambiguous connection between high complexity medical concepts is a critical goal in medical information systems, where the interaction of many factors is mandatory to share the results, and a set of technical and scientific terms with a clear and well-defined meaning should be used.

The last decade witnessed the contribution that the work done in the field of biomedical ontologies when coupled with the data available from healthcare systems (Figure 4) allows disease classification mechanisms among various types of data, thus making it possible for a more refined and dynamically classify of patients. The use of different biomedical ontologies covering different areas of biomedicine results in a substantial improvement in data integration (Haendel, Chute and Robinson, 2018).

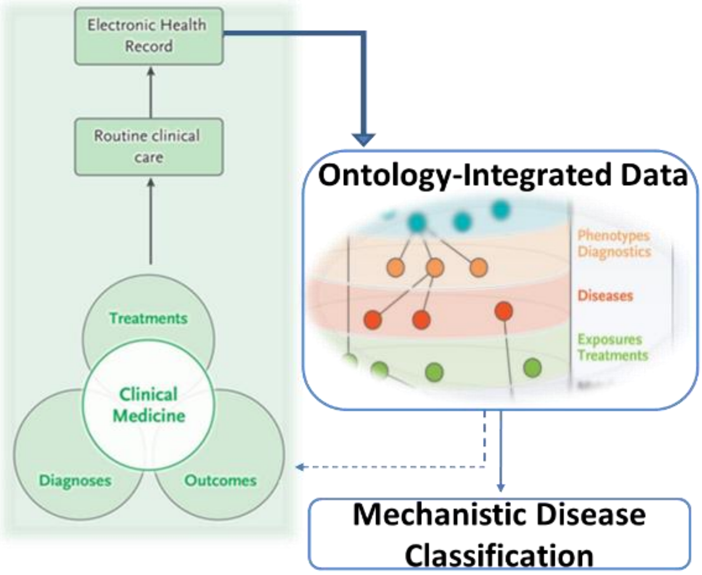


Figure 4: Vision of the relationship between clinical data and biomedical ontologies. Well-structured clinical data can be readily integrated with data originating from research findings using different biomedical ontologies. *Adapted from: Haendel, Melissa & Chute, Christopher & Robinson, Peter. (2018)*

In short, currently in Biomedicine and Medicine fields is increasingly mandatory the integration of controlled biomedical lexical resources in their systems, such as Clinical systems, general medical information systems, medical expert systems, hospital systems, decision support systems, knowledge discovery systems, patient medical records systems, and biomedical text databases, among others (Omid Yousefianzadeh, 2020).

The use of biomedical ontologies has relevant contributions to vocabulary management; data integration, exchange and sharing, knowledge reuse and decision support.

Controlled vocabularies and thesauri are also very relevant in the biomedical domain. They afford less semantics than a true ontology since they typically simply organize terms in a hierarchical structure and provide synonyms and related terms. A relevant resource in this area is the UMLS Metathesaurus, developed by the National Library of Medicine the Unified Medical Language System, it is a system of terminology integration that is constructed by integrating biomedical terms, and it contains several thesauri, controlled vocabularies and even ontologies.

2.2 Semantic Annotation

The main idea of annotation is to enrich the object of study with structurally well-defined associations by associating descriptive and objective descriptions, thus allowing a better understanding of the content and facilitating information extraction (Jovanović and Bagheri, 2017; Larmande and Jibril, 2020). A semantic annotation is an assignment to an entity, present in a given text or data field to a link with its semantic description (Liao *et al.*, 2011; Tchechmedjiev *et al.*, 2018).

In a computer field, semantic Annotation is described as the process of inserting metadata, which are concepts of an ontology (i.e. classes, instances, properties and relations), in Web resources, to assign semantics (Oliveira and Rocha, 2013). Annotating data allows for better search facilities since queries will be based on well-defined concepts described by the ontology of a given domain that it's pretending to search for information instead based only on traditional keywords (Aroyo *et al.*, 2010).

Biomedical semantic annotation has attracted interest from the research community thanks to the many tasks it can support, such as textual semantic management, curation, indexing and facilitated search of data (Jovanović and Bagheri, 2017). By combining the use of ontologies with the annotation process in biomedical repositories, semantic links can be made between different repositories, giving rise to semantic networks of biomedical items, facilitating new scientific discoveries (Jonquet, Shah and Musen, 2009; Jovanović and Bagheri, 2017). This process was initially done manually and therefore was extremely expensive in terms of both resources and time. To tackle these challenges, it was necessary to automate the annotation process (Beasley and Manda, 2018).

To perform an annotation automatically, there are several computational tools, among which the **NCBO Annotator** (Özgür, Hur and He, 2016; Jovanović and Bagheri, 2017; Tchechmedjiev *et al.*, 2018; Perez *et al.*, 2020) **MetaMap** (Stewart, von Maltzahn and Abidi, 2012; Bai *et al.*, 2021), **cTAKES** (Jonquet, Shah and Musen, 2009; Bai *et al.*, 2021), **NOBLE Coder** (Jonquet, Shah and Musen, 2009; Tseytlin *et al.*, 2016) and **ConceptMapper** (Tanenblatt, Coden and Sominsky, 2010; Teng and Verspoor, 2017).

The following scientific articles (Shah *et al.*, 2009; Funk *et al.*, 2014; Galeota and Pelizzola, 2017; Bai *et al.*, 2021) have made several comparisons between some of the previously referenced tools, mainly in performance, which varies depending on the parameters and the objectives defined. In the following section (3.3 - Semantic Annotators) there is a detailed description of these tools.

2.3 Evaluation

To get an idea if the annotation system in use has acceptable performance, one can take advantage of well-known metrics that allows comparing the performance of different systems. In the scientific research context, it is very common to use metrics such as precision (P), recall (R) and F-measure (F1).

Usually, when taking advantage of these metrics for a comparison effect, it is necessary a well-defined corpus also known as a gold standard, from which comparisons are made. These kinds of corpora aim to represent the perfect performance for a given objective as the NCBI Disease corpus, the CRAFT, the Mantra Gold Standard Corpus, and the ShARe among others are good examples of comparison corpora. In this case, a gold standard would be a corpus with a set of annotations made by human annotators, with a high level of expertise in their domain.

In the evaluation of the system, there are 3 possible scenarios, (i)-an annotation that exactly matches the term annotated in the gold standard, (ii)-an annotation that has no match to any term annotated in the gold standard, (iii)-the absence of an annotation of a term that is annotated in the gold standard and (iv)- the absence of an annotation of a term that is annotated in the gold standard. These 3 scenarios are classified as (i)-True positive, (ii)-False positive, (iii)-False negative and (iv)-True negative.

In summary:

- Precision: is the fraction of annotations done by the system that is also present in Gold Standard.

$$Precision = \frac{TP}{TP + FN}$$

- Recall: is the fraction of all terms annotations in Gold Standard that are annotated by the system.

$$Recall = \frac{TP}{TP + FN}$$

- F-measure: is the harmonic mean of precision and recall.

$$F_{measure} = 2 * \frac{P * R}{P + R}$$

2.4 Data Integration

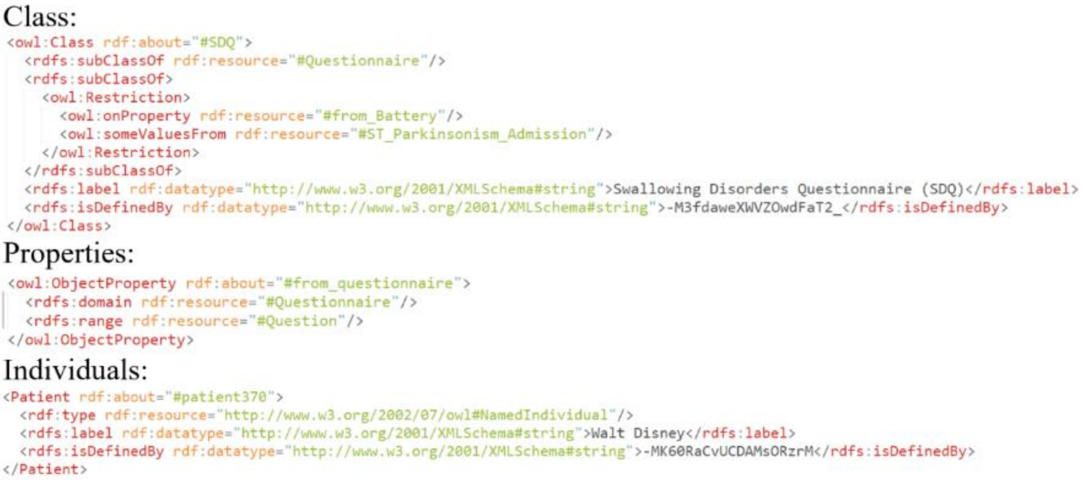
Biomedical data are stored and maintained in various repositories, far exceeding 1500, making the integration process challenging (Jovanović and Bagheri, 2017) but crucial.

Specialization and the consequent increase in the depth of knowledge about a given domain have their importance, however, the crossing of different scientific domains is a broadening of horizons as far as multidisciplinary knowledge is concerned. This transversal knowledge to several scientific areas requires an integration of data from different scientific domains, without ever neglecting the maintenance of detail, uncertainty and of course the context in which the data are inserted (Sioutos *et al.*, 2007; Cheatham and Pesquita, 2017).

This is a process of notable relevance for data generated by healthcare systems, which present a breadth of domains that confers a huge diversity of data and therefore the presence of several heterogeneous entities (Jayaratne *et al.*, 2019; Vidal *et al.*, 2019).

In a general and succinct way, data integration consists in the unification of data that have in common the same semantics but that come from unrelated sources. Semantic data integration solves the heterogeneity problem by employing ontologies to guide the data integration process. A successful integration process reduces data redundancy and the number of queries to be performed and allows the integrated analysis of different data sources.

Many languages and tools could be used in designing ontology for data integration. Regarding used languages, this work only focuses on OWL (Web Ontology Language) which was developed by World Wide Web Consortium (Liao *et al.*, 2011).



Class:

```
<owl:Class rdf:about="#SDQ">
  <rdfs:subClassOf rdf:resource="#Questionnaire"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#from_Battery"/>
      <owl:someValuesFrom rdf:resource="#ST_Parkinsonism_Admission"/>
    </owl:Restriction>
  </rdfs:subClassOf>
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Swallowing Disorders Questionnaire (SDQ)</rdfs:label>
  <rdfs:isDefinedBy rdf:datatype="http://www.w3.org/2001/XMLSchema#string">-M3fdaweXWVZOwdFaT2_</rdfs:isDefinedBy>
</owl:Class>
```

Properties:

```
<owl:ObjectProperty rdf:about="#from_questionnaire">
  <rdfs:domain rdf:resource="#Questionnaire"/>
  <rdfs:range rdf:resource="#Question"/>
</owl:ObjectProperty>
```

Individuals:

```
<Patient rdf:about="#patient370">
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NamedIndividual"/>
  <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Walt Disney</rdfs:label>
  <rdfs:isDefinedBy rdf:datatype="http://www.w3.org/2001/XMLSchema#string">-MK60RaCvUCDAMsORzrM</rdfs:isDefinedBy>
</Patient>
```

Figure 5: Example of how an owl file is structurally organized.

OWL is written using the XML syntax containing three sublanguages, the OWL Lite, OWL DL and OWL Full, and shares several characteristics of RDF (Resource Description Framework) and RDF Schema.

OWL is considered a standard language for ontology representation of the semantic web. Figure 5 shows a simple OWL example from this work, and there are 3 types of components shown: Classes, Properties and individuals.

Chapter 3

Related Work

Table 1 summarizes a set of publications considered relevant to the topic of this thesis.

Table 1: Relevant Related Work.

Publication Title	Brief Description	Authors
The tools and resources for clinical text processing	A survey that brings together a range of available tools, lexical resources, and corpora that are hypothesized to be used in the use of medical context textual data.	(Marovac and Avdic, 2021)
Enhancing cross-lingual semantic annotations using deep network sentence embeddings	Comparative study between the annotation results of the German corpus using the German UMLS and the results of the parallel corpus consisting of medical forms in English and German.	(Lin, Hoffmann and Rahm, 2021)
Cross-lingual semantic annotation of biomedical literature: experiments in Spanish and English	Comparative study between the semantic annotations generated from a Spanish corpus using Spanish UMLS versus the annotations obtained from the parallel corpus consisting of English and Spanish medical forms.	(Perez *et al.*, 2020)
BioBert: a pre-trained biomedical language representation model for biomedical text mining	Paper with the exposition of pre-trained language representation model for the biomedical domain denominated BioBert.	(Lee *et al.*, 2020)
Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies	State of Art revision with an overview of the development and evaluation of NLP algorithms that map clinical texts to ontological concepts. With the final purpose of presenting the level of heterogeneity present in the methodologies used and the establishment of a systematic exposition plan in studies of this kind.	(Kersloot *et al.*, 2020)
Evaluating cross-lingual semantic annotation for medical forms	A Follow-up study by (Lin, Hoffmann and Rahm, 2021) continues the survey but with the integration of deep learning techniques.	(Lin *et al.*, 2020)
SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes	Development of SIFR annotator, a semantic annotator that promises to address the lack of non-English annotation tools.	(Tchechmedjiev *et al.*, 2018)
Bert: pre-training of deep bidirectional transformers for language understanding	Development of BERT, contextualized word representation model with basis on a pre-trained using bidirectional transformers and on a model based on masked language.	(Devlin *et al.*, no date)
Biotea: semantics for PubMed central	Description of the Biotea project that addresses to facilitates the reusing of scientific literature, structuring the information using linked data with standardized web technologies.	(Garcia *et al.*, 2018)

3.1 Medical Forms Annotation

In the last few years, semantic annotation has been used more and more in order to take advantage of the increasing volume of data generated in the most varied areas (Perez *et al.*, 2020). Particularly the implementation of processes of semantic annotation and integration of medical Forms is a process that has been of growing interest (Lin, Hoffmann and Rahm, 2021).

This type of text has great importance in medical research because it allows the integration of knowledge and patterns obtained from clinical data from different patients to extrapolate/predict clinical statuses or trends that may influence the course of action for a given patient (Kersloot *et al.,* 2020).

Typically, medical questions are annotated with several concepts, but it is considered a special case when the whole question itself corresponds to a single concept – Question_as_Concept(QaC); in Lin, Hoffmann and Rahm, 2021 the focus is on these kinds of questions.

Lin *et al*., 2020, annotated a dataset of medical forms and proposed to evaluate the quality of these annotations. Through the results obtained it became evident that the annotation of the corpus of medical forms in German with the ontologies of the German version of UMLS generates very limited results, obtaining a very low annotation retention rate. This low number of annotations generated is a limiting factor since the main objective of the annotation process is to identify as many annotations per question as possible, this being a critical step in obtaining interoperability for the authors' corpora.

Lin *et al*., 2020 also performed a comparative study between the annotation results of the German corpus using the German UMLS and the results of the parallel corpus consisting of medical forms in English and German. Lin *et al*., 2020 also performed a comparative study between the annotation results of the German corpus using the German UMLS and the results of the parallel corpus consisting of medical forms in English and German, authors checked the number of annotations generated on the 37 selected forms, the 37 German forms annotated with the German UMLS and concluded that generate about 3 times less few annotations when compared with the same ontologies but in the English version (UMLS English version). This is evident in the fact that the German UMLS has far fewer concepts than the English version (which has approx. 500 thousand more concepts). It was also observed that in the German UMLS most of the annotations there are of concepts consisting of only one or two words and no QaC are present, indicating that these types of annotations are not present in the German UMLS, unlike the English UMLS.

Lin et al. also annotated the English corpus using a subset of UMLS (it gathers only the ontologies with the most relevance for the context) that was used in the development of a Silver Standard Corpus (SSC) that provides good quality annotations that can be used in the evaluation of cross-lingual semantic annotations. Using this subset, they obtained a 58% improvement in collecting correct annotations, an improvement explained by the fact that in German UMLS lacks several ontologies of crucial relevance, like snomedct and ncit for example. Their results demonstrate that the inclusion of these ontologies is critical for a successful annotation task.

3.2 Cross-lingual Annotation

There is an added challenge in annotating texts/documents in languages like Portuguese, Japanese or French since they have very low coverage in the ontology pool (Lin *et al.*, 2020; Marovac and Avdic, 2021).

In (Lin, Hoffmann and Rahm, 2021) a continuation of previous work (Lin *et al.*, 2020) the authors aim now to identify mainly the annotations of Questions_as_Concept of medical forms in a cross-lingual format using the medical/biomedical concepts present in English UMLS, this is a critical point since QaC are missing of non-English languages.

Both in Lin *et al.*, 2020 and Lin, Hoffmann and Rahm, 2021 the medical forms are originally German, making the annotation process more complex with several variables that can influence the quality of the annotation, one of the main causes is the lack of annotators with satisfactory performance for languages other than English.

The annotation process of a non-English document, also known as Cross-lingual annotation (Lin *et al.*, 2020; Perez *et al.*, 2020) involves the assignment of concepts from English ontologies to text segments of non-English documents, but only after applying machine translation tools on the original corpus, thus achieving an annotation with much higher coverage of terms. In summary, the cross-lingual annotation approach provides 2 advantages, the use of the UMLS of the English version which is much more comprehensive than in any other language and the fact that several annotators can be applied since most of the developed annotators target English corpora.

To evaluate the quality of cross-lingual annotation, it is necessary to build an English SSC that allows comparison with automatic cross-lingual annotations. Unlike previous works that only used automatic annotation tools to generate an initial annotation set, the authors of Lin et al., 2020 added manual verification of the generated annotations to the pipeline. This action has an impact on the accuracy of the SSC construction since the verification of automatic annotations is done using the 2-vote-agreement protocol that sometimes can have a problem in verifying the annotation, more specifically in the case where 2 of the 3 e.g. annotators annotate the same way a target text, this supposedly indicates that the annotation is correct, however, the annotation may be incorrect and this protocol (2-vote-agreement) only demonstrates that the annotators work similarly, the manual verification avoids this problem.

3.3 Semantic Annotators

There are two types of semantic annotators, the human annotator, and the automatic annotator that can also be divided into two categories: term-to-concept matching approach and approach based on machine learning (ML). All have naturally the same goal, the assignment of the ontological concepts presents in a given text fragment. The human annotator does the annotation manually, generally, more than one human annotator is involved, firstly they decide individually and then discuss in a group which is the most appropriate concept to assign. In case there are several concepts assigned to the same text segment, the one with the most accurate description (based on the definition, synonyms and semantic types) is chosen; If there are several concepts that perfectly fit the target text, all those concepts are kept (different concepts but with adequate semantic types); finally, if the complete sentence is linked to a UMLS concept, then the recommendation is to assign that concept (Question_as_Concept) and the concepts that are assigned to the annotated segments of that text fragment.

There are several tools used for semantic annotation which can suffer from different types of limitations:

- Speed: because the huge datasets take a lot of time in processing.

- Language specificity: the majority of tools in English, makes the application of semantic annotation in other languages impossible.

- Document type: some annotators support documents as input and with this possibility problems such as annotating different document formats or not supporting a specific format arise

- Text variation: The fact that to exist different types of biomedical texts and variations in texts, e.g. biomedical vs. clinical texts can be a challenge (Jovanović and Bagheri, 2017). For example, biomedical literature is full of acronyms, on another side, the clinical texts have many variations with the locals' dialects.

- Entity disambiguation: Biomedical entities without enough context in texts, which makes their disambiguation difficult.

In the following, the most popular semantic annotation tools are described.

NCBO Annotator[4] is a web service that allows easily the linking of a biomedical target text with the knowledge contained in the ontologies present in the BioPortal and UMLS methasaurus repositories; for this reason has considerable importance in the biomedical field, being an annotator used by the National Center for Biomedical Ontology (NCBO) to index biomedical resources and improve information retrieval and data integration in the biomedical domain(Tchechmedjiev *et al.*, 2018; Perez *et al.*, 2020). Unlike most annotators, it uses a method to associate concepts, instead of looking for the concept with the best match score for a given context. The main special characteristic of this annotator is the fact that it takes advantage of BioPortal which allows suitable real-time processing, despite doesn't support the disambiguation of terms. BioPortal is an enormous online repository of approximately 900 biomedical ontologies, created and currently maintained by the NCBO.

The NCBO Annotator annotates textual data with ontology terms from UMLS and BioPortal having, therefore, a wide range of ontologies available, of the set of tools presented, is the only one capable of associating a concept with several other related concepts, instead of finding a single concept with the best association score and associate only that one (Jonquet, Shah and Musen, 2009), this is relevant to perform annotations that achieve greater coverage, since it is much more valuable to get an annotation set that can cover an entire sentence with several concepts instead of a single concept that maps the sentence, the more concepts the annotator maps, the more coverage it provides.

NCBO annotator is an annotator that can be divided into two stages, one that uses the MGrep term-to-concept matching tool and another that retrieves sets of annotations that are later expanded using various methods of semantic matching.

How does it work?

● The user gives a document (e.g free text) as input to a concept recognition tool jointly with a dictionary. This dictionary is also known as lexicon is a set of strings that identifies ontology concepts. The construction of this dictionary is based on accessing ontologies and grouping all concept names or other string forms (synonyms, labels) which syntactically identify concepts.

● NCBO Annotator takes advantage of Mgrep2 to recognize concepts using string matching on the dictionary. This first set generated of direct annotations is used as input

[4] http://bioportal.bioontology.org

for the semantic expansion components, which expand the annotations collected from the previous step using the knowledge represented in one or more ontologies.

Over the last few years, the NCBO BioPortal has been progressively improving, adopting domain-independent and open source semantic web technologies. This evolution allows any researcher to take advantage of the virtual NCBO implemented in its methodology with the necessary adjustments to take full advantage of the NCBO features in a personalized way.

The constant evolution in this field allows for the constant emergence of new annotation tools, which always bring novelties and new annotation possibilities. this is the case with the **SIFR annotator**.

As is well known, the English language largely dominates the scientific community, however, there are more and more biomedical data that are originally in languages other than English. A paradigmatic example is the language adopted by clinicians when generating data, which is usually done in the local language and not in the clinician's native language. In this project(Tchechmedjiev *et al.*, 2018), SIFR annotator, the authors address this problem, where the language adopted is French in which there is a considerable gap in the volume of terminologies and ontologies making it difficult to treat these data in a facilitated way, this is a problem that can be extended to languages other than English.

With this work, the authors intend to find a solution that takes advantage of the huge amount of biomedical data that is produced in French, such as electronic health records. For this, they have developed the SIFR BioPortal platform, through the Semantic Indexing of French Biomedical Data Resources (SIFR) project, this open-source platform has integrated the French ontologies and terminologies that are present in NCBO.

In this way, this work focused on a platform of various services for searching, browsing, mapping hosting, mapping generation, the possibility of describing and editing semantically rich metadata, versioning, visualization, recommendations, and community feedback. This platform aims to facilitate the processing of texts and clinical notes in French, using French ontologies and terminologies and taking advantage of an annotation website.

The NCBO BioPortal was crucial in the development of the SIFR annotator, serving as a template from which customizations and improvements were made to be able to cope with French texts, but in this case, instead of serving mostly English ontologies/terminologies, the focus will be on French ontologies. So these ontology repositories have as main function to enable the annotation with ontologies of French biomedical texts or notes, the way the selection of these annotations is done can vary, for example through semantic groups or types.

The SFIR annotator uses 30 terminologies and ontologies from the SIFR BioPortal web platform, so this platform is a local instantiation of the NCBO technology but in this case, targeted at data in Frances that in addition to identifying entities also performs entity linking by mapping explicit ontology classes to entities.

cTAKES is a modular system of combined components rule-based and machine learning techniques with the main goal of information extraction from clinical data (Savova *et al.*, 2010; Jovanović and Bagheri, 2017). Is another annotator, which components are mainly trained for the clinical domain, developing a relevant pool of rich linguistic and semantic annotations. These annotations are the baseline for several methods and modules for semantic processing of clinical free-text at a high level.

Currently "cTAKES" is compound by the following components/annotators (Savova *et al.*, 2010):

● The sentence boundary detector extends OpenNLP’s supervised ME sentence detector tool. Also allows the prediction of whether a period, question mark, or exclamation mark is at the end of a sentence.

● Tokenizer: consists of a component that splits the sentence internal text stream on the space and punctuation; and in another component that is context-dependent tokenizer, in other words, merge tokens to create date, fraction, measurement, person title, range, roman numeral, and time tokens by applying rules for each of these types.

● Normalizer: is a wrapper around a component of the "SPECIALIST" Lexical Tools (Browne *et al.*, 2003) called “norm”. This component allows assigning a representation for each word in the original text that is normalized respecting the lexical properties and can map multiple mentions from the same word that do not have the same string representations in the input data.

● Part-of-speech (POS) tagger and Shallow parser: As Normalizer are wrappers around OpenNLP’s modules for these tasks. Savova *et al.*, 2010 study provides at that moment a new supervised ME model trained on manually annotated clinical data. POS tagging is very useful for automatically analysing human speech data and forms the backbone of NLP engines (translation apps for example). Based on the output from the shallow parser, the algorithm finds all noun phrases, which become the look-up window.

● Negation annotators: In charge of the implementation of the NegEx algorithm, this approach is pattern-based for finding words/phrases indicating negation near named entity mentions.

● Named entity recognition (NER) annotator: Implements a terminology-agnostic dictionary lookup algorithm within a noun-phrase look-up window. Each named entity is mapped to a concept from the terminology with the use of the dictionary lookup. Named entities refer to an element of the documents/text which belong to a particular class from a set of predefined specific classes. This cTAKES's component does not resolve ambiguities that result from identifying multiple terms in the same text span.

There are three approaches of NER, rule-based, machine learning-based and hybrid approaches.

● Status annotator: The status annotator uses a similar approach to the negation annotator but in this case aims to find relevant words/phrases that indicate the status of a named entity.

ConceptMapper is a purpose dictionary lookup tool, highly configurable, flexible and accurate, implemented as an open-source UIMA component, as part of an NLP system. (Tanenblatt, Coden and Sominsky, 2010; Jovanović and Bagheri, 2017)

Unlike the other annotators explained here, conceptMapper was not developed with the main aim on the biomedical domain but is rather generic and configurable enough to apply to any domain (Tanenblatt, Coden and Sominsky, 2010; Jovanović and Bagheri, 2017). For the ConceptMapper’s operation, the tokenizer is the only thing necessary to have been run before ConceptMapper, though a sentence detector is also usually useful.

ConceptMapper was developed as a flexible tool that can provide accurate mappings of unstructured text into named entities, as specified by dictionaries vocabulary controlled. The time performance of this tool is very high which allows a real-time result with million entry dictionaries. Takes advantage of a dictionary that stores various possible variants per each entry and then connects them to the same concept, this allows handling of a good variety of ways a concept can be mentioned in the input text (synonyms and distinct word forms) (Tanenblatt, Coden and Sominsky, 2010). The individual entries in a dictionary may consist of multiple tokens that could potentially be assigned to a non-contiguous text.

In conceptMapper lookups are token-based, and limited to a specific context, normally a sentence, but is highly configurable for any context needed, such as a noun phrase or other NLP-based concepts.

Features that can be reconfigured (Tanenblatt, Coden and Sominsky, 2010):

- Type of annotations that are created and the features are associated with those annotations.
- Processing of input document tokens
- Lookup strategy
- Finally, there are a set of post-processing filters, and an interface to create new filters, this gives the possibility of over-generating results during the lookup phase, and consequently reduce the result set according to particular rules.

Finally, according to (Perera, Dehmer and Emmert-Streib, 2020), in the last years, Deep learning boosted the evolution of natural language processing (NLP) leading to advances in machine reading on a large scale, biological analysis, and database curation. Therefore, combining the approach using NPL tools with Named Entity Recognition (NER) models is a way with appetising potential.

NER is mainly referred in the state of art as entity identification and this system has been growing in importance in the biomedical domain since the amount of biological data generated in digital text files has been increasing. An equally important feature of NER is entity extraction, which consists of a subtask of information extraction that has the main goal of summarising knowledge into expressive forms for management and understanding, finding entities and categorizing target text. This gives great support to decision-making. (Lin *et al.*, 2020; Perera, Dehmer and Emmert-Streib, 2020).

NER and Relation Detection (RD), allows the research and identification of interactions between separate concepts that have some point of connection, e.g the interaction between symptom and disease or gene and disease. On a large scale, these interactions can be translated into networks to summarize details on a given biomedical or clinical task, and then able for data management and further in an easier way. (Yadav and Bethard, 2019; Perera, Dehmer and Emmert-Streib, 2020).

In the biomedical field, there have been some BioNLP tools that are NLP tools but specialised for biomedical data, and NLP tools are used to identify entities and relations in the text. Biomedical entities are denoted by groups into classes (genes, symptoms, diseases etc). (Yadav and Bethard, 2019)

The importance of NER systems was already discussed above but for an additional feature that allows the classification beyond identification, it is feasible to use the

"Named Entity Recognition and Classification" (NERC), which have the same logic as NER but have added value since in addition to the identification of given entities this system have their classification into standard or normalized terms. In Summary, both systems are almost equal varying in this detail. (Perera, Dehmer and Emmert-Streib, 2020)

An aimed NER system at Biomedical is BioNER which is frequently used in the state of the art as a starting point in text mining tasks such as summarizing, text/document classification, associations between biological entities and biological networks. Therefore, in a medical context is a good hypothesis to resolve the same text-related tasks.

3.4 Machine Translation

The Lin *et al*., 2020 authors throughout their study highlighted and reinforced the use of ontology-based annotations to improve interoperability and the quality of data integration in both healthcare and biomedical research domains. This study also compares and evaluates the performance of different annotators on English and German corporations, and between different translation tools.

The use of translation tools is central to the success of the annotation process of non-English documents/texts, allowing the use of the English version of the UMLS in the annotation. A poor translation compromises the success of the annotation and consequently of the semantic integration. However, nowadays it is still practically impossible that the application of automatic translation tools allows results close to the annotations on corpora originally in English, since the translations contribute to deviation from the original forms, consequently the translated forms instead of being similar to the original forms they are similar to paraphrase of original forms (Lin, Hoffmann and Rahm, 2021).

For a choice of viable translation tools, the authors of Lin *et al*., 2020 first collected a random set of 50 questions from the original corpus (German) and then translated all questions into English using 5 translation tools: DeepL, Microsoft translator, google translate, yandex and moses. After these translations were obtained, a manual analysis of the results was made. For each tool output, 1 point was given for each question with the best translation; choosing the two best translators: DeepL and Microsoft Translator.

The comparison between the annotation results with the different chosen translation tools revealed DeepL produces the best annotation result outperforming Microsoft translator, with higher values in the recall and precision metrics and consequently better F-measure. This indicates that using only a small number of translated samples (50 in this case) it is possible to assess the viability of a translation tool.

The integration of machine translation tools in the workflow of this study allowed increasing the retention of correct annotations by about 70% concerning the annotation on the original corpus. This is a very relevant fact for this study, which includes medical forms in Portuguese that will have to be translated to obtain the most acceptable annotation results possible.

The authors of Lin *et al.*, 2020 also showed that by using 2-vote-agreement it is possible to combine the results coming from 2 different translation tools to improve the scores of the 3 metrics under evaluation, the accuracy, recall and F-measure. In total, the rates of annotations retained when the combination between the two tools was used improved in relation to DeepL alone by 10.3% (from 58.0% to 68.3%) and in relation to Microsoft translator alone by 14.2% (from 54.1% to 68.3%). However, (Lin, Hoffmann and Rahm, 2021) uses a different strategy, different results were obtained where the tool that works best was Google Translator.

Chapter 4

Design and Implementation

This chapter overviews the methodology designed to build a semantic model for a database that stores standardized clinical questionnaires and, through the integration of external ontologies, support the semantic annotation of different questionnaires and evaluation scales and their subsequent semantic integration. It also details the

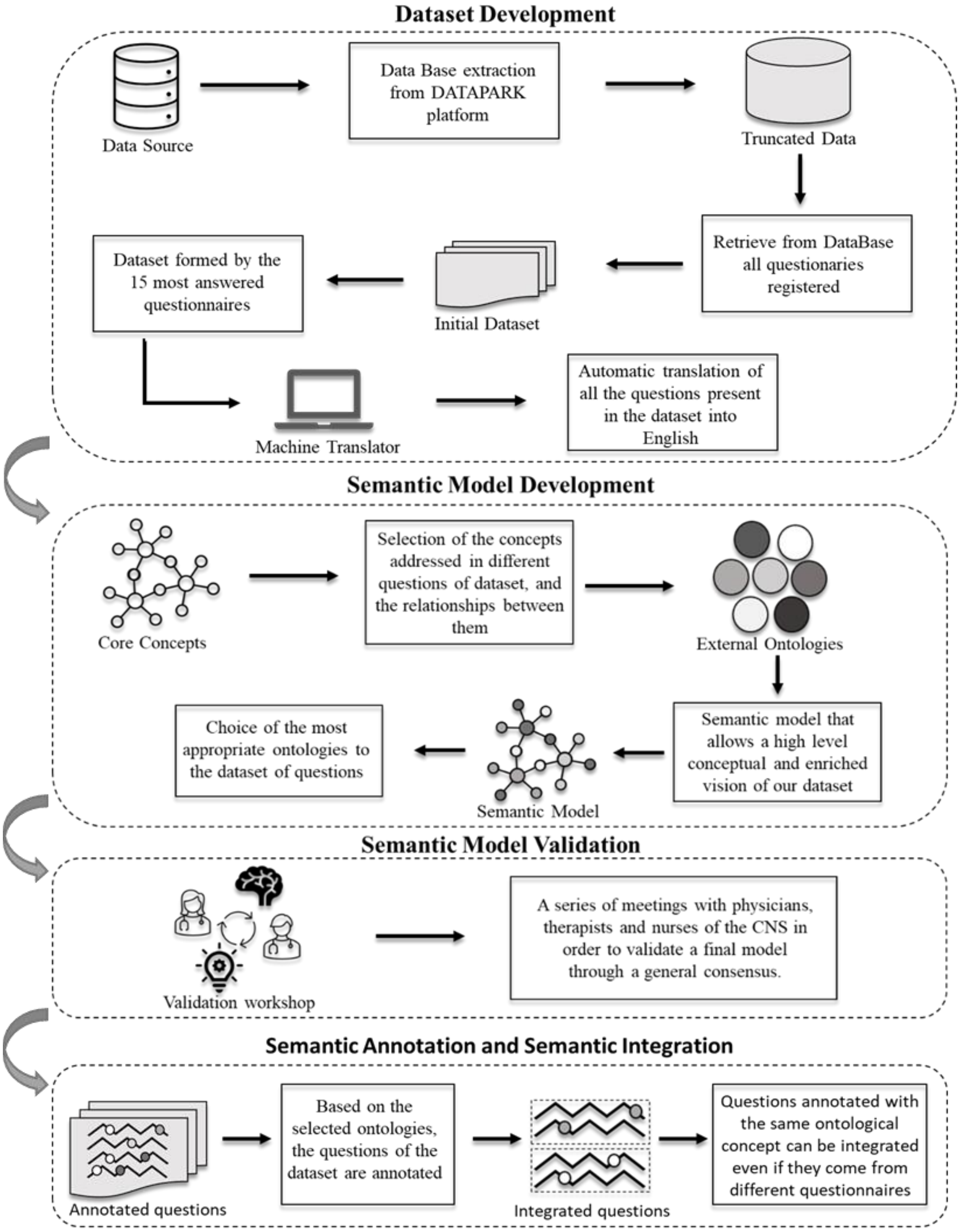

Dataset Development
Data Source
Data Base extraction from DATAPARK platform
Truncated Data
Retrieve from DataBase all questionaries registered
Initial Dataset
Dataset formed by the 15 most answered questionnaires
Machine Translator
Automatic translation of all the questions present in the dataset into English
Semantic Model Development
Core Concepts
Selection of the concepts addressed in different questions of dataset, and the relationships between them
External Ontologies
Semantic model that allows a high level conceptual and enriched vision of our dataset
Semantic Model
Choice of the most appropriate ontologies to the dataset of questions
Semantic Model Validation
Validation workshop
A series of meetings with physicians, therapists and nurses of the CNS in order to validate a final model through a general consensus.
Semantic Annotation and Semantic Integration
Annotated questions
Based on the selected ontologies, the questions of the dataset are annotated
Integrated questions
Questions annotated with the same ontological concept can be integrated even if they come from different questionnaires

Figure 6: Overview of the whole methodology built and followed in this thesis.

implementation process. represents a diagram of the overview of the methodology developed and followed.

Figure 6 is represented the overview of the entire pipeline. This methodology was organized and based on 4 main pillars: Dataset Development, Semantic Model Development, Semantic Validation and finally Semantic Annotation and integration. In Dataset development are all the transformations and processing that the data undergoes since it is extracted until the formation of the final dataset. In the development of the semantic model begins the process that relates the data to its semantic content, and ontologies belonging, this model has an added value since it was developed in partnership with a team of physicians and therapists of the CNS.

Finally, the semantic annotation of each of the questions presents in the Dataset allowing then the integration of them.

The methodology also includes a semi-automated or manual validation of several of the steps, to ensure validation of the methodology before it is replicated for all questionnaires in the database in future work.

4.1 Dataset Development

The DataPark platform is supported by a document-oriented database. The document store has a hierarchical organisation where at the top are the most comprehensive clinical categories, within each category, where are the tests batteries (sets of related questionnaires or evaluation scales), within each battery, there are questionnaires or evaluation scales, each questionnaire is populated with several questions, and each question has multiple possible answers. When a questionnaire is applied to a patient, the system stores information including a patient identifier, the date and time, and the specific answer to a specific question item.

To get an idea of the volume of data present in this database was carried out a quantitative survey of all the data categories as the below table shows.

Table 2: Volume of Data in Dataset.

Number of Batteries	Number of Questionnaries	Number of Questions
28	98	>4000

Since a semi-automated or manual validation of several of the steps is required, the working dataset is a subset of the whole database. The relevance of each questionnaire was evaluated based on the number of answers that were registered in the database. The final selection included the 5 most answered clinical questionnaires related to parkinsonism, the 5 most answered clinical questionnaires related to stroke and finally the 5 most answered clinical questionnaires not related to a specific pathology, but globally. The goal was to achieve sufficient coverage of different disorders while still maintaining the workload of semi-automated or manual evaluation manageable. After the extraction of these 15 questionnaires, the preliminary DataSet was built. These data were extracted in JSON format, for further processing. A subset of the data in JSON format is shown in Appendix A.

4.2 Automated translation of the questionnaires

The DataPark database contains both questionnaires in their original version (English) and questionnaires translated into Portuguese, which naturally increases the heterogeneity of the data. Therefore, to standardise the data in a single language, technologies were integrated that allow the detection and subsequent automatic translation into a previously defined language.

To choose the translation tool to be integrated into the system, a comparative evaluation was run. This study consisted in identifying certain questionnaires that, although they are in the database in Portuguese, are originally written in English. The identification of these questionnaires allowed the creation of a corpus of the same questions in English (since they are available in external sources) and in Portuguese (from the DataPark system). Therefore, was used those same questions but in Portuguese, the language in which they are represented in the database and translated them using the translation tools that are in comparison. All the outputs from these translation processes were compared with the English question corpus and were evaluated their performances using the metrics Recall, Precision and F1. This evaluation can be seen in detail in the Results section, where it’s observed that the tool which shows the highest reliability is the Microsoft translator.

4.3 Machine translation

The program built iteratively runs through the Dataset of 15 questionnaires and each question is introduced as input in a python function that takes advantage of Microsoft's language detection API and gives as output an acronym that represents the language in which that question is, for example, Portuguese-"pt", Spanish-"es", English-"en" etc. After that, this output is used as a variable that will serve to filter the next step, which can follow two paths: if that acronym corresponds to the English language ('en') then the program does nothing, allocating the question just as it entered the iteration loop. On the other hand, if the acronym is different from 'en' then the program takes advantage of another python function but now to do the translation. The translation function receives 2 inputs, the question in the non-English language and the descriptive acronym of that language.

Now the translation function that uses the translation API from Microsoft, translates the input question from the source language (passed in the acronym) to English. Once the translation is done, the function gives the output in the form of an already translated question, which is used to take the place of the same question, but in the other language.

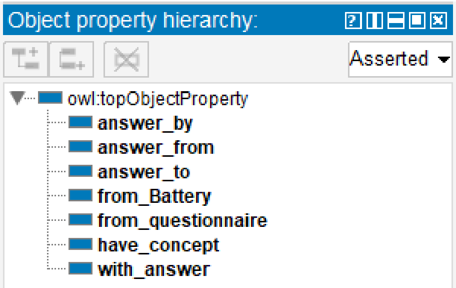
Once all the iteration is finished, the dataset is returned with all the questions that had other languages translated into English, in other words, the final dataset is completely in the English language.

4.4 Semantic model design and integration

The construction of the initial semantic model was divided into 2 steps: the initial model design based on the contents of the questionnaires and scales and the alignment of the model with existing selected ontologies. The alignment of the basic model supports the semantic enrichment of the descriptions of the questions, providing a broader semantic scope.

4.4.1 Initial semantic model design

The first iteration of the initial semantic model was designed by me, through a process of analysing each questionnaire and manually extracting the high-level concepts referred to by them. The extracted concepts were organized in a hierarchy, and relationships were established between them. The Semantic Editor tool Protégé was used to support the implementation of the model in OWL. The main concepts were modelled as OWL Classes and the following types of relationships (modelled as OWL Object Properties) were also created as shown in Figure 7.



- owl:topObjectProperty
  - answer_by
  - answer_from
  - answer_to
  - from_Battery
  - from_questionnaire
  - have_concept
  - with_answer

Figure 7: OWL Object Properties.

For example, the identification of the concepts Social life, Mood, Anxiousness, Sadness and Angriness allows identifying potential relationships of hierarchy between them. Sadness, Anxiousness and Angriness can be modelled as subclasses of Mood. There are also other kinds of relationships between the concepts, for instance, Sadness influences Social Life.

The initial model provided an overview of the main topics covered by the questionnaires.

4.4.2 Integration with external ontologies

The Integration with external ontologies step is extremely important since it is here that the set of ontologies to be used in the annotation is chosen. Therefore, to have the set that best fits the type of data present in the DataPark platform was used a dataset of 15 questionnaires was evaluated using the Ontology Recommender service of the NCBO Bioportal. This service receives biomedical texts and suggests which ontology or ontologies are most appropriate based on the terms referenced in the provided texts.(Martínez-romero *et al.*, 2017)

Therefore, using this service allows the assessment of the content of all the questions in the 15 questionnaires and thus describe which ontologies best match the content inserted in the questionnaires.

Let's take as an example a Dataset composed of the following questions from the GDS questionnaire:

Do you often **feel bored**? Are you generally in a **good mood**? Are you **afraid** something **bad** is going to happen to him? Do you usually **feel happy**? Do you often **feel forsawed**? Would you rather **stay at home** instead of going out and **doing new things**? Do you **feel** you have more **problems** with your **memory** than most other **people**? Do you think it's **good to be alive**? Do you **feel useless**? Do you **feel** too **much energy**? Do you **feel** that your situation is **desperate**? Do you **feel** that most **people's** situation is **better** than yours?

In bold are the terms that ncbo recommender identifies and according to these terms, the service will find the ontology or the set of ontologies that best covers the whole dataset, that is, ncbo recommender evaluates the dataset as a whole and not question by question. For this particular dataset, the results with the highest coverage are:

- Set composed by a single ontology - NCIT (52.2%);
- Set composed by two ontologies -> NCIT and LOINC (68.7%);
- Set composed by three ontologies -> NCIT and LOINC and OCHV (81.1%).

However, if when this analysis question by question is done, the results obtained are different, for example,for the question "Do you feel useless? " the results are:

- Set composed of a single ontology -> GALEN (66.7%);
- Set composed of two ontologies -> GALEN and OCHV (100%), i.e. it is easier to find a set of ontologies that achieves a high coverage score since there are far fewer terms to be mapped.

For this reason, the NCBO Bioportal service meets the identified needs and therefore will be used within the pipeline using the BioPortal API.

Through this API it is possible to integrate a code block in a program that allows to map the dataset, i.e. automatically evaluate which ontologies have higher coverage of collected data and consequently to the model. The analysis measures that the ncbo recommender provides are coverage score, acceptance score, detail score, specialization score, number of annotations and a final score. The most important measures were coverage score and final score because they allow choosing the ontologies that give greater security in the correct mapping of data.

The ncbo recommender allows to obtain coverage analysis based on just one ontology or more than one, this allows studying which approach obtains more coverage score in the selected metrics. In other words, instead of using one or another ontology to perform the annotation, it may be more profitable to use a simultaneous set of ontologies.

4.5 Semantic model validation

The model validation is a long process to achieve a high level of quality, coverage and consensus since it does not depend only on one or two people, but on a multidisciplinary team, composed of doctors, therapists, and nurses from CNS.

A preliminary step in the validation process included three LASIGE researchers from the Faculty of Sciences of Ulisboa who have already been involved in the DataPark project. This preparatory step had as objective the revision and improvement of the preliminary semantic model. Although the elements involved in this initial validation are not clinical experts, they have a working knowledge of the DataPark platform and the data it stores, which enabled the production of an improved version of the semantic model.

The improved model was presented to the director of the Campus Neurológico Senior, a neurology expert, who then identified a multidisciplinary team at CNS who would work directly with me to study the model presented and make the necessary alterations to better reflect their vision.

This teamwork, called "validation workshop", had the collaboration of seven teams from the CNS and their Clinical Director, who represented the respective clinical areas, Nutrition, Psychology, Speech Therapy, Occupational Therapy, Physiotherapy, Neuropsychology and Nursing. The methodology of this validation workshop was applied individually to all these clinical areas.

Therefore, the validation workshop was organised according to the following points:

1. Firstly, it is necessary to assess which concepts (or terms) are more relevant to describe the pertinent information of the different questionnaires and scales most commonly used in each clinical area. Naturally, this relevance will change according to the vision obtained from the different areas of action. Within each area, and under the coordination of the clinical manager of that area, a selection is made of the 10 most relevant concepts in the area and of the 5 concepts that are considered most important, but which are not exclusive to their area of specialisation (global concepts).

2. The next step involves meetings with the responsible persons in different areas. These meetings will aim to organise and enrich the concepts identified, defining a hierarchy of concepts and more informative relationships between them, with a focus on establishing conceptual relationships between concepts from the same area and between concepts from different areas.

3. Finally, after all the individual meetings, a final meeting is held with all the areas in which the final validation of the model presented is carried out, this validation is obtained through the consensus of all those present at the validation workshop.

4.6 Semantic Annotation of Questionnaires

The annotation of the questionnaires, that is, their questions, is a crucial point in this methodology since the quality of the annotation will have considerable repercussions on the semantic data integration, the last step of this methodology.

To perform the semantic annotation, there is a need to find the classes in the selected ontologies that a developed python program receives as input a file containing a sub data set composed by the hierarchy Diagnosis area, tests battery, questionnaires, questions, and answer possibilities. The program developed takes advantage of several useful libraries such as ElasticSearch[5], fuzzywuzzy, owlready2, nltk and the API of the NCBO annotator tool and the API of the ElasticSearch.

The annotation process as explained in Figure 8 is divided into 2 sub-processes, firstly the textual processing of the question and then the annotation of the question.

[5] https://www.elastic.co/

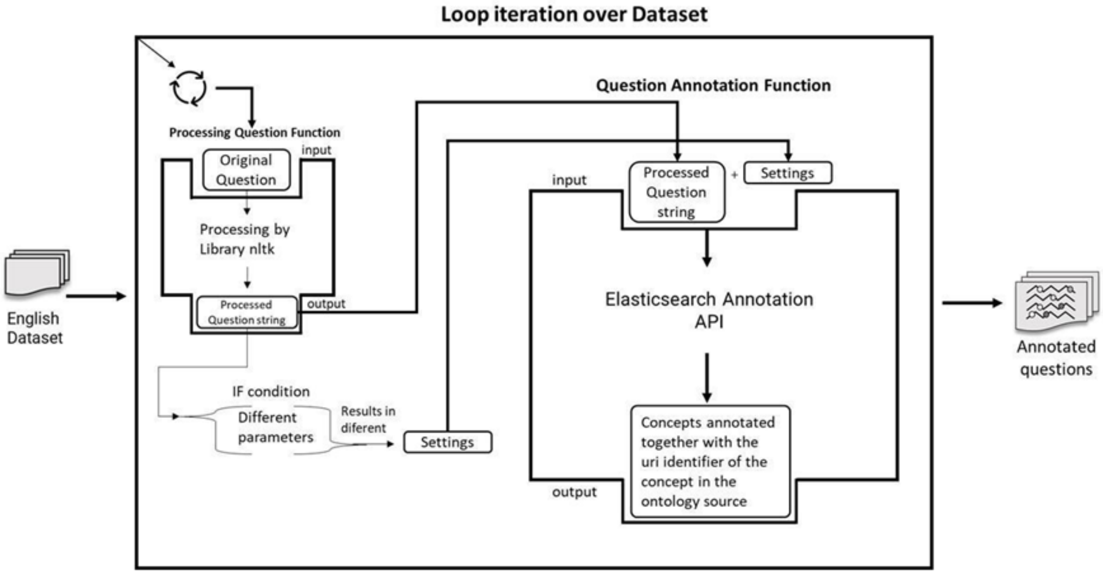

Loop iteration over Dataset
Processing Question Function
Original Question
input
Processing by Library nltk
Processed Question string
output
IF condition
Different parameters
Results in diferent
Settings
Question Annotation Function
input
Processed Question string
+
Settings
Elasticsearch Annotation API
Concepts annotated together with the uri identifier of the concept in the ontology source
output
English Dataset
Annotated questions

Figure 8: Semantic Annotation chart flow.

The textual processing step aims at increasing the accuracy of the annotation of non-QaC terms, by first removing stopwords (i.e., words with low relevance) such as "the", "a", "an", "in" among many others. This step also checks the length of the sentence that makes up the question. Naturally, the performance of the annotation process varies as the characteristics of the sentence vary, so in this process, specific settings are defined that will have an impact on the annotation process. These settings are the base of the annotation approach to the type of sentence to be annotated. For shorter sentences, the program takes advantage of ElasticSearch settings through match_query and query_string query, for longer sentences it uses match_phrase_query, multi_match_query and match_phrase_prefix_query.

Then enters the second sub-process, that of the annotation itself:

In the first iteration, the Question Annotation step was conducted using the NCBO annotator API, however, we found some limitations, including:

- The NCBO annotator API has a very small word limit of about 500 words when the sub-dataset has more than 10000 words;
- The processing time is rather long (3 times longer than ElasticSearch);
- There is a difficulty in the disambiguation of annotations for polysemic terms (which decreases precision)

- The NCBO annotator API struggles with the use of abbreviations (which decreases recall);

- Finally, NCBO annotator API provides the very little possibility of parameterizing the process and adapting it to the specificities of the dataset.

To overcome these limitations, an implementation of a straightforward semantic annotation tool based on ElasticSearch was performed. ElasticSearch is a Lucene-based open-source search engine that is used to find ambiguous questions and for this purpose have an indexing engine that provides both superior query performance and rich query syntax to face off a massive number of queries for a large volume of data. Although it offers the possibility of dealing with a massive volume of data this distributed search engine have high effectiveness, stability, and scalability. As this service works by a RESTful server, i.e., the communication with it is through its REST API. This tool makes use of indexing and lookup functions provided by ElasticSearch to perform annotation.

This newly developed tool performs annotation with an approach that varies depending on the length of the sentence and the objective that was initially outlined. Four cascading objectives were defined:

1. Search for annotations as concepts (i.e., Question as Concept)
2. Annotations composed of 3 words
3. Annotations composed of 2 words
4. Finally, annotations are composed of a single word.

The final result of the annotation is the composition of all the processes. Since ElasticSearch works based on indexes that store scores, it is possible to sort the output from the highest score to the lowest one, or just scores higher than a given value, or even choose the TOP 3 of the highest scores, for instance. This cascading approach is geared towards finding the most specific annotations possible, targeting the longest terms first.

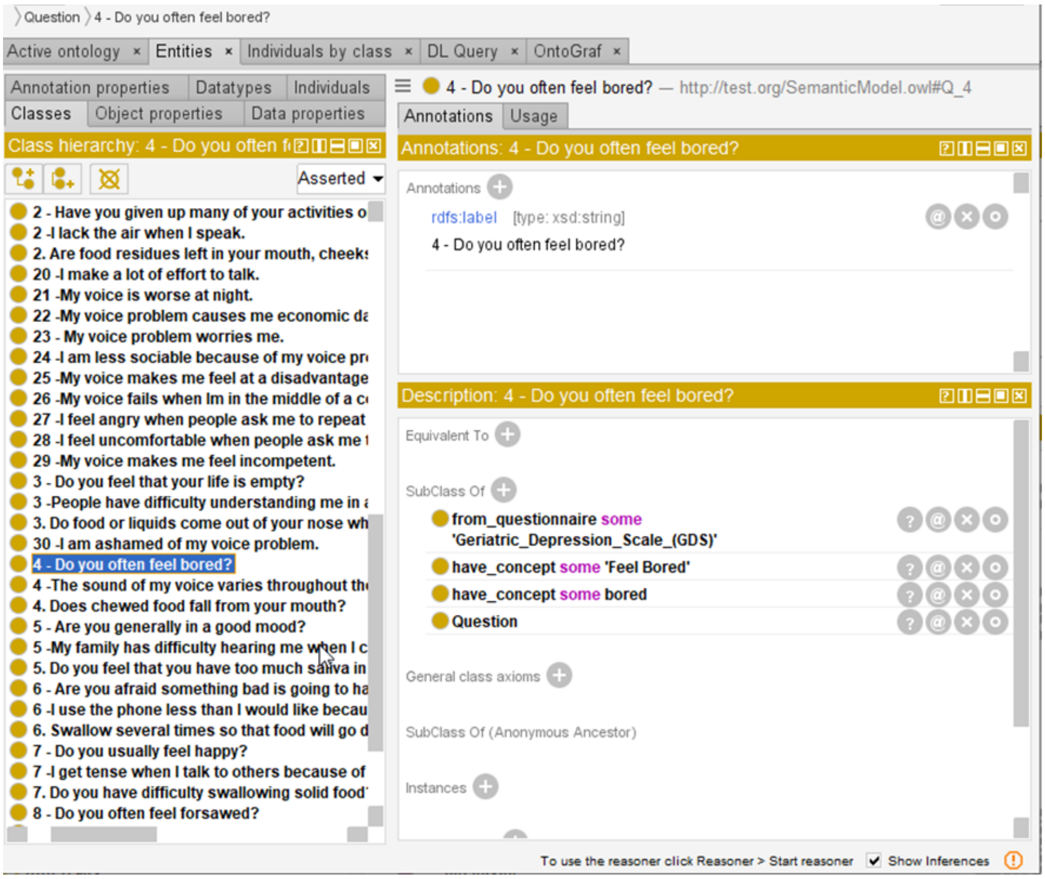
In this program, I opted to join the score given by ElasticSearch to the score given by an algorithm from the FuzzyWuzzy library that measures the strength of similarity between strings, through Levenshtein Distance to calculate the differences between sequences or terms. After several empirical tests, I gave 80% weight to the score of ElasticSearch and 20% to fuzzywuzzy, was with these percentages that I get a better performance of annotation.

After the annotation is generated an output that is always composed of the ontological term and its identifying URI, that works as an ID of that concept belongs to an ontology. In the end, all the questions associated are mapped with ontology concepts. This association is encoded in an OWL file created through the owlready2 library, which allows the creation of mappings between the questions and the concepts that they mention that are present in the selected ontologies. Each question will keep, therefore, besides an association to its battery of tests, diagnostic area, and possibilities of answering one or more associations to ontological concepts that have been mapped in the question by the developed program.

4.7 Semantic Integration of Questionnaires

After the semantic annotation of the questions is accomplished, the next step is the semantic integration of actual responses and evaluations for patients. With this, it becomes possible to evaluate a given patient in relation to a given concept that may be covered by several clinical areas and questionnaires.

Figure 9 shows how questions are modelled, the Question, "Do you often feel bored?" is a subclass of the Geriatric Depression Scale (that is a subclass of Questionnaire). Questions have established Data Properties to their controlled set of answers. For instance, "Do you often feel bored?" has a relation "has_answer" with a given number of answer options: yes, not, almost always, never. Also has a relation "has_concept" with the ontology concept "bored".



〉Question 〉4 - Do you often feel bored?

Active ontology × | Entities × | Individuals by class × | DL Query × | OntoGraf ×

Annotation properties | Datatypes | Individuals
Classes | Object properties | Data properties

Class hierarchy: 4 - Do you often f

Asserted

- 2 - Have you given up many of your activities o
- 2 -I lack the air when I speak.
- 2. Are food residues left in your mouth, cheek
- 20 -I make a lot of effort to talk.
- 21 -My voice is worse at night.
- 22 -My voice problem causes me economic da
- 23 - My voice problem worries me.
- 24 -I am less sociable because of my voice pr
- 25 -My voice makes me feel at a disadvantage
- 26 -My voice fails when Im in the middle of a c
- 27 -I feel angry when people ask me to repeat
- 28 -I feel uncomfortable when people ask me
- 29 -My voice makes me feel incompetent.
- 3 - Do you feel that your life is empty?
- 3 -People have difficulty understanding me in
- 3. Do food or liquids come out of your nose wh
- 30 -I am ashamed of my voice problem.
- 4 - Do you often feel bored?
- 4 -The sound of my voice varies throughout th
- 4. Does chewed food fall from your mouth?
- 5 - Are you generally in a good mood?
- 5 -My family has difficulty hearing me when I c
- 5. Do you feel that you have too much saliva in
- 6 - Are you afraid something bad is going to ha
- 6 -I use the phone less than I would like becau
- 6. Swallow several times so that food will go d
- 7 - Do you usually feel happy?
- 7 -I get tense when I talk to others because of
- 7. Do you have difficulty swallowing solid food'
- 8 - Do you often feel forsawed?

4 - Do you often feel bored? — http://test.org/SemanticModel.owl#Q_4

Annotations | Usage

Annotations: 4 - Do you often feel bored?

Annotations
rdfs:label [type: xsd:string]
4 - Do you often feel bored?

Description: 4 - Do you often feel bored?

Equivalent To

SubClass Of
- from_questionnaire some 'Geriatric_Depression_Scale_(GDS)'
- have_concept some 'Feel Bored'
- have_concept some bored
- Question

General class axioms

SubClass Of (Anonymous Ancestor)

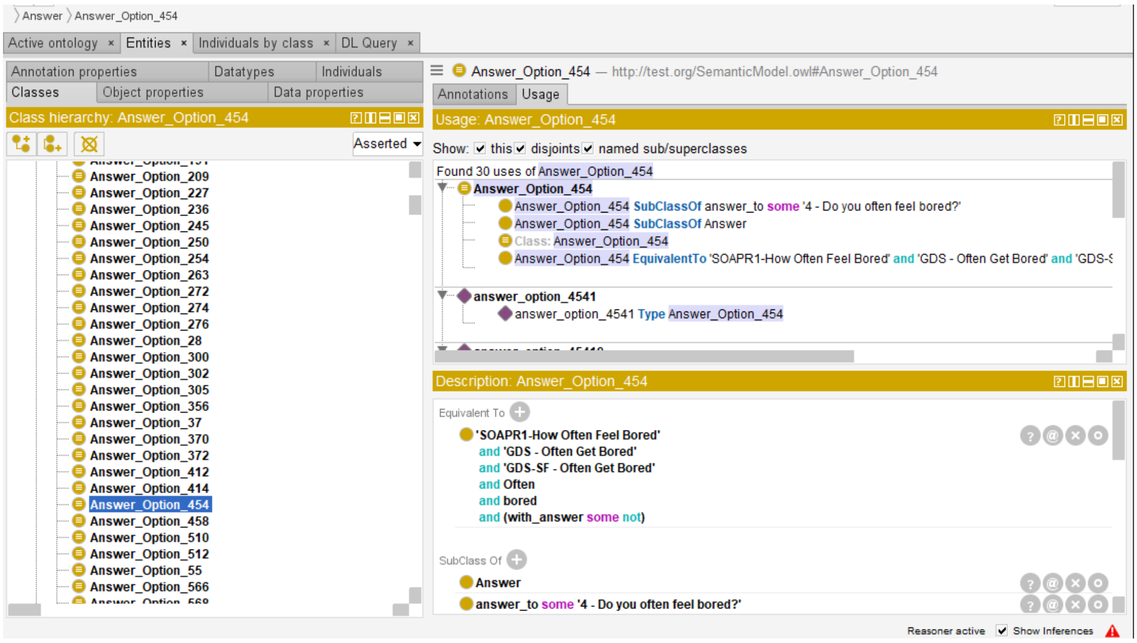
Instances

To use the reasoner click Reasoner > Start reasoner ☑ Show Inferences

Figure 9: Demonstration of how is modelled the questions and their concepts in protégé.

The 2 figures below show how the modelling and integration of the questionnaire data, its questions, the concepts in them with the answers given by the patients is done.

In the 1st figure, Figure 10 shows the class "Answer" that stores the possible answer option, that is, the question concepts and a given "status". Whenever a patient has registered an answer to this question with that "status" an individual is created.



Answer_Option_209
Answer_Option_227
Answer_Option_236
Answer_Option_245
Answer_Option_250
Answer_Option_254
Answer_Option_263
Answer_Option_272
Answer_Option_274
Answer_Option_276
Answer_Option_28
Answer_Option_300
Answer_Option_302
Answer_Option_305
Answer_Option_356
Answer_Option_37
Answer_Option_370
Answer_Option_372
Answer_Option_412
Answer_Option_414
Answer_Option_454
Answer_Option_458
Answer_Option_510
Answer_Option_512
Answer_Option_55
Answer_Option_566

Answer_Option_454 — http://test.org/SemanticModel.owl#Answer_Option_454

Show: ☑ this ☑ disjoints ☑ named sub/superclasses

Found 30 uses of Answer_Option_454

Answer_Option_454
Answer_Option_454 SubClassOf answer_to some '4 - Do you often feel bored?'
Answer_Option_454 SubClassOf Answer
Class: Answer_Option_454
Answer_Option_454 EquivalentTo 'SOAPR1-How Often Feel Bored' and 'GDS - Often Get Bored' and 'GDS-S

answer_option_4541
answer_option_4541 Type Answer_Option_454

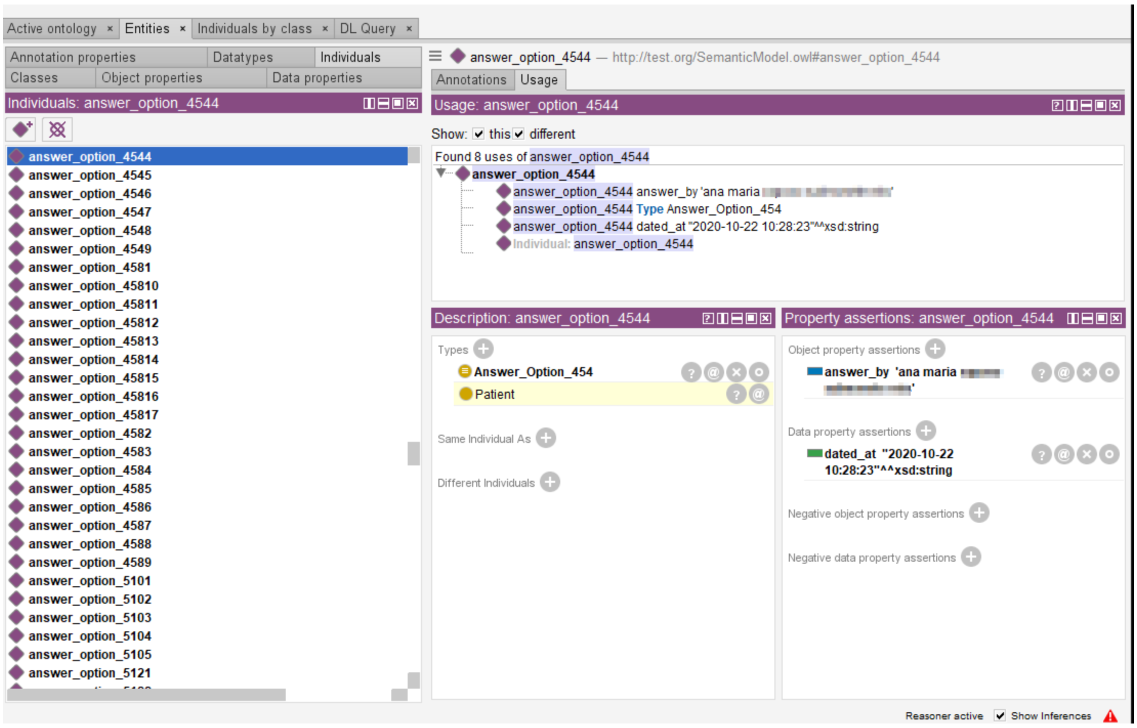
Equivalent To
'SOAPR1-How Often Feel Bored'
and 'GDS - Often Get Bored'
and 'GDS-SF - Often Get Bored'
and Often
and bored
and (with_answer some not)

SubClass Of
Answer
answer_to some '4 - Do you often feel bored?'

Figure 10: Demonstration of how answer option appears in protégé.

The individual that corresponds to a patient response record is saved with the prefix "answer_option" followed by an incremental number, and this instance is related with patient individual and associated with a data property that stores the date of the patient answer.

The 2nd figure, Figure 11 shows how the relationship between the questions in the questionnaire and the patient data works. This record("answer_option ") " answer_option _4544" is associated with a Patient individual “ana maria” and a Date, is also related to the answer option " Answer_Option_454" that corresponds to a Status of "not" given by the data property with_answer and the concepts "Do you often get bored", "Feel", and "Bored".



answer_option_4544 — http://test.org/SemanticModel.owl#answer_option_4544

Individuals: answer_option_4544

answer_option_4544
answer_option_4545
answer_option_4546
answer_option_4547
answer_option_4548
answer_option_4549
answer_option_4581
answer_option_45810
answer_option_45811
answer_option_45812
answer_option_45813
answer_option_45814
answer_option_45815
answer_option_45816
answer_option_45817
answer_option_4582
answer_option_4583
answer_option_4584
answer_option_4585
answer_option_4586
answer_option_4587
answer_option_4588
answer_option_4589
answer_option_5101
answer_option_5102
answer_option_5103
answer_option_5104
answer_option_5105
answer_option_5121

Usage: answer_option_4544

Show: ☑ this ☑ different

Found 8 uses of answer_option_4544

answer_option_4544
- answer_option_4544 answer_by 'ana maria [redacted]'
- answer_option_4544 Type Answer_Option_454
- answer_option_4544 dated_at "2020-10-22 10:28:23"^^xsd:string
- Individual: answer_option_4544

Description: answer_option_4544

Types
- Answer_Option_454
- Patient

Same Individual As

Different Individuals

Property assertions: answer_option_4544

Object property assertions
- answer_by 'ana maria [redacted]'

Data property assertions
- dated_at "2020-10-22 10:28:23"^^xsd:string

Negative object property assertions

Negative data property assertions

Reasoner active ☑ Show Inferences

Figure 11: Demonstration of how answers individuals appears in protégé.

In short, the ontology concepts annotate a given question. In this way it´s can do the whole flow from the battery of tests, through the questionnaire, the question and the concepts annotated in it, and integrate all this information with patient data.

Thus, a concept-oriented view of the patient's health status and its evolution over time.

Chapter 5

Results and Discussion

This chapter presents the results achieved throughout the explained methodology in Chapter 4, with a discussion of the features, performance and feasibility of the developed pipeline.

5.1 Machine Translation

The original dataset does not only contain data in English, there is also a large fraction in Portuguese - the local language of the clinic from which the data was collected.

For the reason evoked above, I conducted a study parallel to the construction of the semantic annotation/integration pipeline, in which I investigated which would be the best machine translation tool to use. Based on the literature review I identified two candidates to be incorporated in the pipeline, Microsoft Translator - MT and Google Translator - GT.

This study was organised in 3 phases:

1) The first phase consisted in identifying 3 questionnaires from the dataset that was in Portuguese but had their original version in English in the state-of-art. The identification of these questionnaires with original versions in English allowed me to create a Gold Standard corpus composed of 60 questions from 3 different questionnaires:

i) Geriatric Depression Scale – GDS,

ii) Swallowing Disturbance Questionaire – SDQ,

iii) Voice Handicap Index – VHI

2) The second phase consists in the translation from Portuguese to English of these 60 questions present in the dataset. To achieve that, the pipeline was tested with the MT API and the GT API, proceed with the translation and establish outputs for both tools.

3) The third and last phase is to evaluate the outputs generated by the two tools and identify the one with the best performance. This evaluation is done by comparing the questions translated into English with the questions in the corpus which are originally in English. This comparison can be made following the procedure of two different systems, System A and System B. System A is a more rigid system in which only consider a translation well done if the words match perfectly. System B is more permeable to

changes in the translated words as long as the semantics remains unchanged such as verb tenses or plural words. For these reasons, I gave more weight to the results generated with this System.

After carrying out the 3 phases the qualitative results achieved with the MT tool using the A and B system and the same for the GT tool are obtained. To translate these results from qualitative to quantitative the true positives and negatives as well as the false positives and negatives need to be computed. It was then possible to calculate precision, recall and F1. The final results can be seen in tables 3 and 4 and Figure12.

Table 3: System B Machine Translator performance evaluation results.

	Precision		Recall		F-Measure	
	MT	Google	MT	Google	MT	Google
GDS	**0.8188**	0.8176	**0.7503**	0.7492	**0.1934**	0.1929
SQS	0.7433	**0.7531**	**0.7120**	0.7000	**0.1807**	0.1799
VHI	**0.7817**	0.7635	**0.8117**	0.7829	**0.1980**	0.1913
Average	0.7813	0.7781	0.7580	0.7440	0.1907	0.1880

The table above (Table 3) shows the results of system B where it can be verified that in the 3 questionnaires the difference in the use of one or another translation tool is not very high. Using MT the precision is bigger with a difference of 0.003%, the recall presents a superior value of 0.014 and finally, the F-measure is 0.009% superior when compared with the GT result. Therefore, using this system, the MT presents a better performance even if for a small difference.

Table 4: System A Machine Translator performance evaluation results

	Precision		Recall		F-Measure	
	MT	Google	MT	Google	MT	Google
GDS	0.7932	**0.7999**	0.7270	**0.7314**	0.1873	**0.1885**
SQS	0.6361	**0.6960**	0.6093	**0.6467**	0.1547	**0.1665**
VHI	**0.7623**	0.7352	**0.7932**	0.7564	**0.1932**	0.1845
Average	0.7305	0.7437	0.7098	0.7115	0.1784	0.1798

The results presented in table 4, show that the use of system A presents better results than using GT. The precision with more 0.0132%, recall with a higher value of 0.0017% and the F-measure with more than 0.007%.

In both the results generated with systems A and B the differences between the two tools are practically negligible.

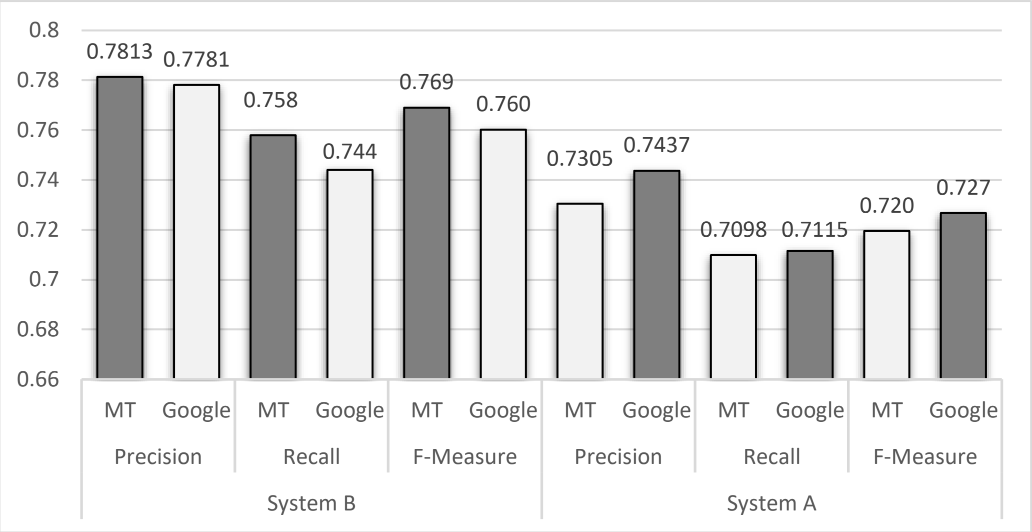


Figure 12: System A and B combined performance results.

Figure 12 shows a view of the previous results together, where it can be seen that with the use of system B as the means of evaluation results on 78.13% accuracy, 75.80% recall and an F-measure of 0.769 for the MT results which give some confidence in the use of Microsoft translator in the pipeline. The results also show that the use of GT would also be successful. The performance of the translation is evaluated more faithfully if system B is used since when 2 sentences say the same without using the same words, while system A evaluates such a situation as a mistranslation, system B accepts and classifies it as a correct translation.

5.2 Ontology Selection

The study was carried out under the gathered DataSet of Questions to reach the Set of ontologies that will serve as a basis for the semantic annotations of the questionnaires, a correct choice of ontology set is crucial for a successful annotation process and consequently a smooth integration process.

Since the dataset contains a lot of text, it was not possible to study the entire dataset. Therefore, this study includes 500 words at a time until the entire dataset is completed. This justifies repeating the same set of ontologies throughout the results. The principle is that the set with the highest coverage rate should be chosen.

The results with a single ontology are shown in Table 5. With only one ontology the results (a TOP10 was made) were low, with the maximum coverage reached being 66.5% with NCIT and 61.8% with OCHV, which is to be expected since one single ontology cannot be expected to provide adequate coverage of a data set with a considerable level of heterogeneity in domains as is the case.

Table 5: Results for sets of single ontology.

Ontology	Final Score	Coverage Score	Acceptance Score	Detail Score	Specialization Score	Annotations
NCIT	0.679	**0.665**	0.858	0.786	0.447	201
NCIT	0.625	**0.622**	0.858	0.797	0.229	220
OCHV	0.462	**0.618**	0.277	0.256	0.278	196
NCIT	0.624	**0.617**	0.858	0.837	0.205	156
OCHV	0.435	**0.577**	0.277	0.253	0.256	208
OCHV	0.433	**0.572**	0.277	0.268	0.245	167
OCHV	0.46	**0.571**	0.277	0.25	0.444	192
NCIT	0.599	**0.558**	0.858	0.782	0.305	157
OCHV	0.429	**0.552**	0.277	0.282	0.275	158
NCIT	0.594	**0.548**	0.858	0.809	0.282	202

Table 6 presents the results obtained with ontology pairs. In this situation the coverage values of the dataset increased significantly, the minimum coverage in the obtained data (a TOP10 was made) was higher than the maximum value of the previous study, 72.6% being the new maximum of 81.5% with the NCIT and OCHV. Note that 5 ontologies appear in the results of the set, NCIT - 10/10, OCHV 6/10, RCD 2/10 and LOINC and SNOMED 1/10 each.

Table 6: Results for sets of two ontologies.

Ontology	Final Score	Coverage Score	Acceptance Score	Detail Score	Specialization Score	Annotations
NCIT, OCHV	0.673	**0.815**	0.649	0.603	0.247	258
NCIT, OCHV	0.7	**0.783**	0.706	0.646	0.446	227
NCIT, LOINC	0.7	**0.752**	0.852	0.639	0.421	202
NCIT, OCHV	0.654	**0.749**	0.72	0.699	0.197	198
NCIT, SNOMEDCT	0.708	**0.743**	0.873	0.73	0.39	203
NCIT, OCHV	0.64	**0.743**	0.654	0.614	0.273	259
NCIT, RCD	0.694	**0.733**	0.858	0.685	0.393	203
NCIT, OCHV	0.627	**0.731**	0.645	0.638	0.218	208
NCIT, OCHV	0.623	**0.727**	0.652	0.626	0.212	216
NCIT, RCD	0.658	**0.726**	0.858	0.668	0.2	231

When the study was done for 3 ontologies in the set, the results in Table 7 (a top 20 was made) show again the dominance of NCIT and SNOMED in the ontology sets, the maximum coverage value increases to 84.6% with NCIT, LOINC and OCHV. Note that NCIT is again in all sets 20/20, OCHV in 19/20, SNOMED 3/20 and Loinc 5/20, the other ontologies are more sporadically present.

Thus evaluated data the set that was chosen for further annotation consists of NCIT, LOINC, SNOMEDCT and the OCHV.

Table 7: Results for sets of three ontologies.

Ontology	Final Score	Coverage Score	Acceptance Score	Detail Score	Specialization Score	Annotations
NCIT, LOINC, OCHV	0.725	**0.846**	0.746	0.56	0.427	227
NCIT, SNOMEDCT, OCHV	0.701	**0.837**	0.75	0.635	0.217	258
NCIT, OCHV, RCD	0.683	**0.837**	0.666	0.583	0.237	254
NCIT, OCHV, LOINC	0.681	**0.836**	0.657	0.577	0.244	254
NCIT, OCHV, NIFSTD	0.677	**0.826**	0.641	0.602	0.244	260
NCIT, OCHV, UPHENO	0.676	**0.826**	0.631	0.61	0.239	258
NCIT, OCHV, HUPSON	0.677	**0.825**	0.641	0.601	0.244	259
NCIT, OCHV, IOBC	0.676	**0.824**	0.641	0.603	0.242	257
NCIT, OCHV, MEDDRA	0.678	**0.823**	0.664	0.598	0.24	258
NCIT, OCHV, OBA	0.675	**0.821**	0.643	0.604	0.245	258
NCIT, OCHV, ENVO	0.675	**0.819**	0.647	0.603	0.246	260
NCIT, OCHV, FMA	0.674	**0.817**	0.651	0.602	0.245	257
NCIT, OCHV, MESH	0.674	**0.816**	0.65	0.604	0.246	258
NCIT, OCHV, UBERON	0.674	**0.816**	0.649	0.604	0.246	258
NCIT, SNOMEDCT, OCHV	0.727	**0.814**	0.794	0.666	0.401	227
NCIT, OCHV, LOINC	0.673	**0.814**	0.684	0.548	0.273	194
NCIT, LOINC, SNOMEDCT	0.726	**0.813**	0.864	0.611	0.384	201
NCIT, OCHV, RCD	0.711	**0.81**	0.723	0.62	0.428	226
NCIT, LOINC, OCHV	0.685	**0.81**	0.753	0.657	0.189	201
NCIT, LOINC, OCHV	0.668	**0.806**	0.709	0.578	0.212	221

5.3 Semantic Model

The semantic model developed was initially discussed individually with the CNS departments of Nursing, Physical Therapy, Speech Therapy, Occupational Therapy, Psychology, Neurology, and Nutrition, and later unanimously approved by all department heads together with the Clinical Director.

The model gathers a Total of 204 medical concepts present in the ontology set, of which 7 are super-classes that aggregate 196 sub-classes. In total, these concepts have 1089 relations among themselves.

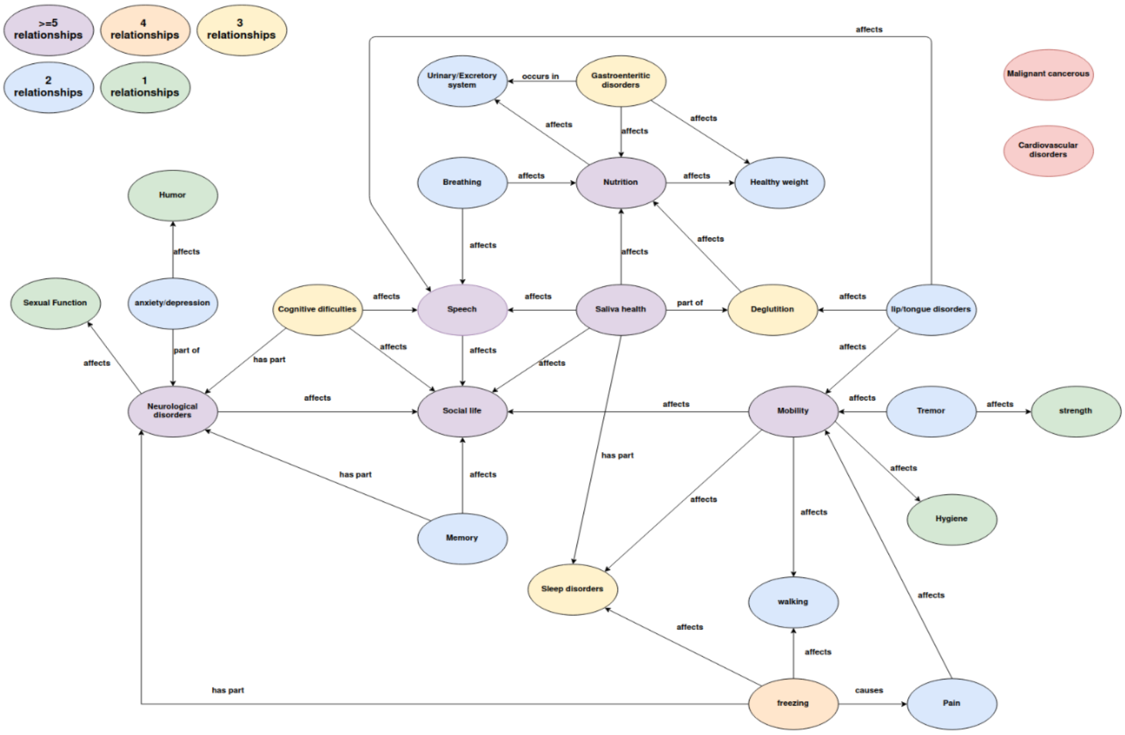
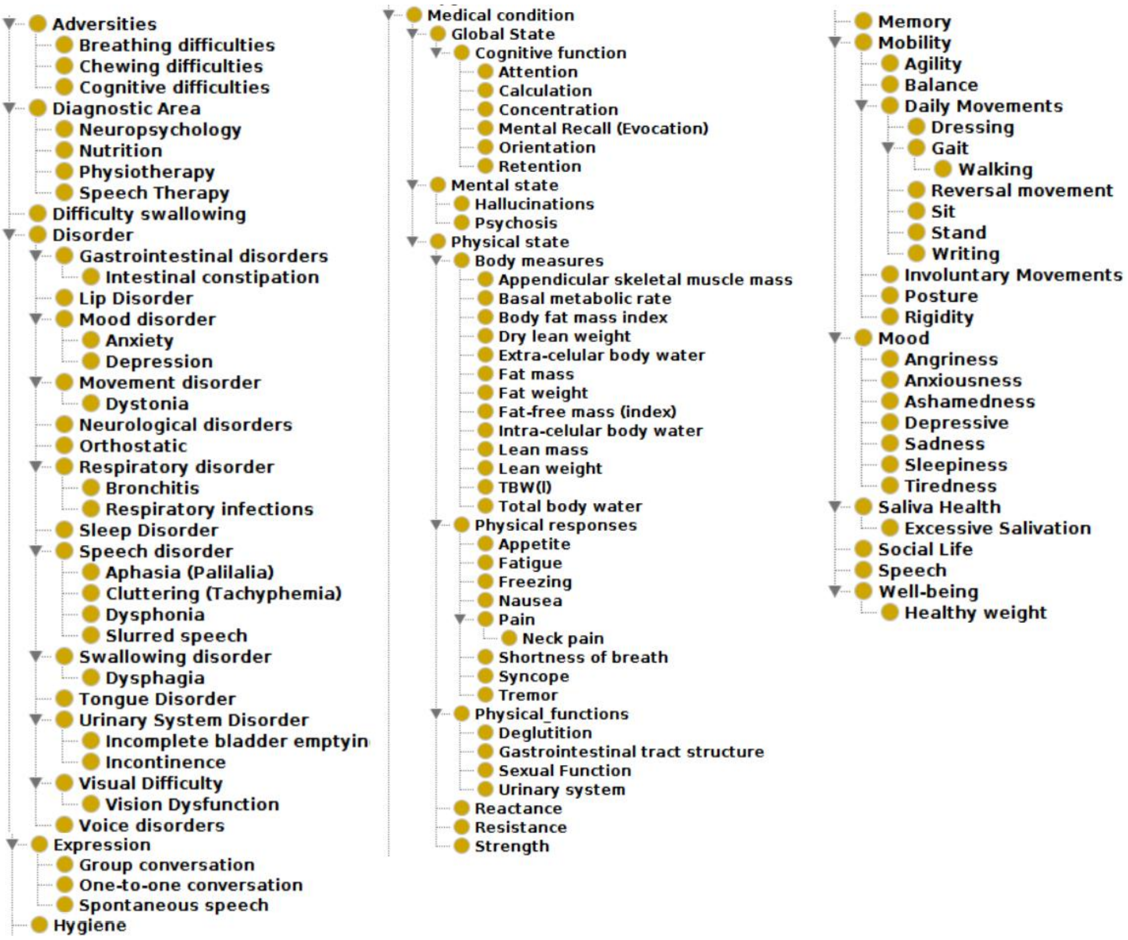


Figure 13: First draft of the semantic model.

Figure 13 shows the first draft of the semantic model. The number of concepts extracted in this first phase was reduced and was obtained through a high-level survey of the content of the questions in the sub-dataset questionnaires. In total, 25 concepts were extracted, from which 36 relations were identified. This first version was created manually, using the draw.io[6] tool that allows users to create all types of diagrams.

[6] https://app.diagrams.net/

In this version (Figure 14) and the following ones, the development of the model has been done using a more appropriate tool for this purpose, protégé[7]. This tool allows the user to create and manipulate hierarchy and dependency relationships between terminologies that are (or are not) supported in ontologies. In this way, it can not only continue the construction of the model with a deeper search for concepts present in the questions of the sub-dataset, but it can also integrate these concepts of the questions with terminologies present in the ontologies. This version presents improvements in the depth that it reaches when compared to version 1(Figure 13). In more detail, it has 118 concepts and 455 relations, which represents an increase of 472% of concepts and 1263% of relations compared to the previous version. This version has a maximum depth of 4 subclasses, i.e. it can go down to a level of granularity that allows going down 4 levels from the parent concept to the lowest hierarchy descendant concept.



- Adversities
  - Breathing difficulties
  - Chewing difficulties
  - Cognitive difficulties
- Diagnostic Area
  - Neuropsychology
  - Nutrition
  - Physiotherapy
  - Speech Therapy
- Difficulty swallowing
- Disorder
  - Gastrointestinal disorders
    - Intestinal constipation
  - Lip Disorder
  - Mood disorder
    - Anxiety
    - Depression
  - Movement disorder
    - Dystonia
  - Neurological disorders
  - Orthostatic
  - Respiratory disorder
    - Bronchitis
    - Respiratory infections
  - Sleep Disorder
  - Speech disorder
    - Aphasia (Palilalia)
    - Cluttering (Tachyphemia)
    - Dysphonia
    - Slurred speech
  - Swallowing disorder
    - Dysphagia
  - Tongue Disorder
  - Urinary System Disorder
    - Incomplete bladder emptyin
    - Incontinence
  - Visual Difficulty
    - Vision Dysfunction
  - Voice disorders
- Expression
  - Group conversation
  - One-to-one conversation
  - Spontaneous speech
- Hygiene
- Medical condition
  - Global State
    - Cognitive function
      - Attention
      - Calculation
      - Concentration
      - Mental Recall (Evocation)
      - Orientation
      - Retention
  - Mental state
    - Hallucinations
    - Psychosis
  - Physical state
    - Body measures
      - Appendicular skeletal muscle mass
      - Basal metabolic rate
      - Body fat mass index
      - Dry lean weight
      - Extra-celular body water
      - Fat mass
      - Fat weight
      - Fat-free mass (index)
      - Intra-celular body water
      - Lean mass
      - Lean weight
      - TBW(I)
      - Total body water
    - Physical responses
      - Appetite
      - Fatigue
      - Freezing
      - Nausea
      - Pain
        - Neck pain
      - Shortness of breath
      - Syncope
      - Tremor
    - Physical_functions
      - Deglutition
      - Gastrointestinal tract structure
      - Sexual Function
      - Urinary system
    - Reactance
    - Resistance
    - Strength
- Memory
- Mobility
  - Agility
  - Balance
  - Daily Movements
    - Dressing
    - Gait
      - Walking
    - Reversal movement
    - Sit
    - Stand
    - Writing
  - Involuntary Movements
  - Posture
  - Rigidity
- Mood
  - Angriness
  - Anxiousness
  - Ashamedness
  - Depressive
  - Sadness
  - Sleepiness
  - Tiredness
- Saliva Health
  - Excessive Salivation
- Social Life
- Speech
- Well-being
  - Healthy weight

Figure 14: First version made in Protégé.

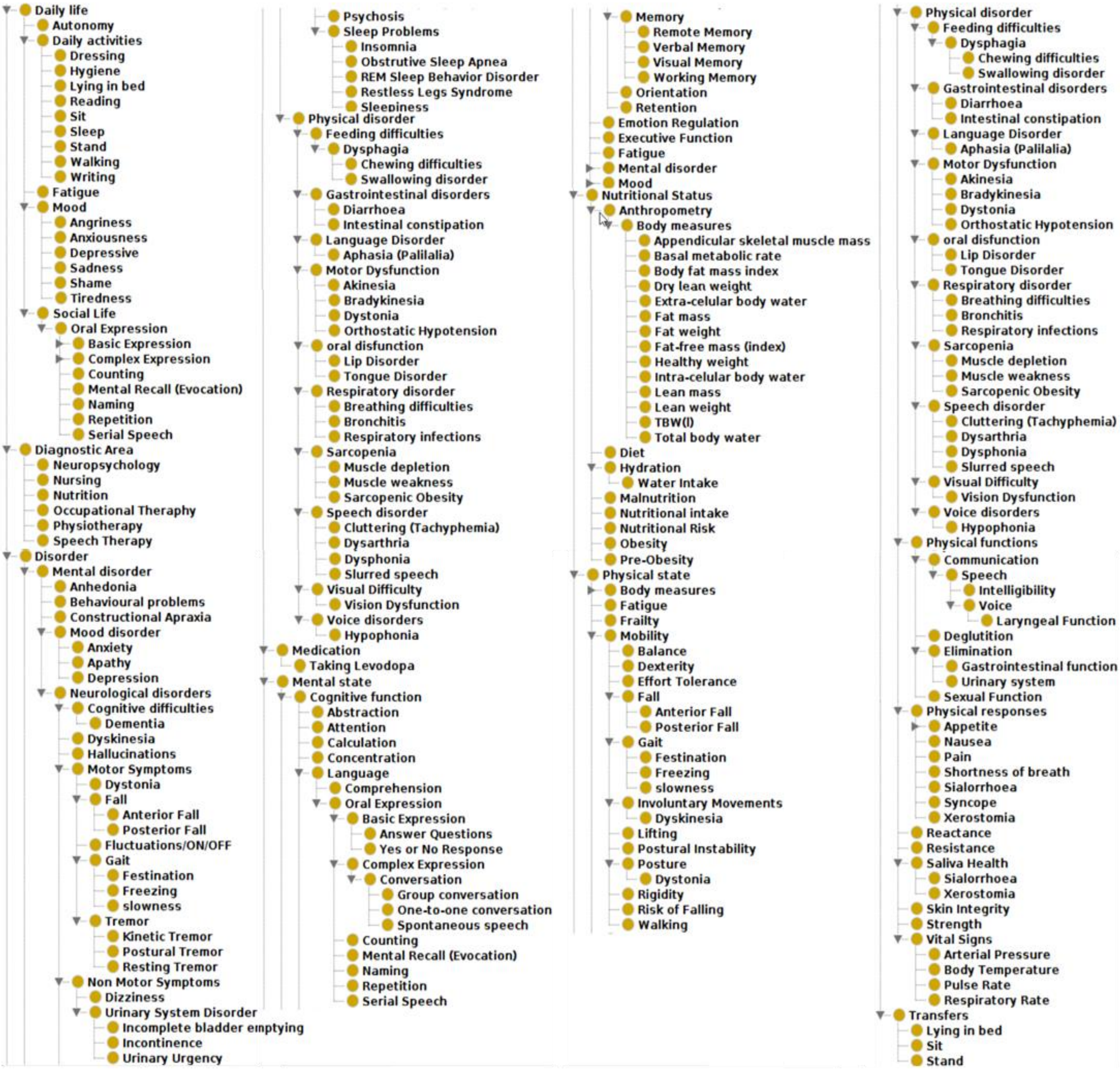
The improvements observed in this phase of the semantic model are observed both in the number of identified concepts and in the relevance that they have in the clinical context, as well as in the relationships established between concepts. These

[7] https://protege.stanford.edu/

improvements were due to the meetings with LASIGE researchers and DataPark developers who contributed with their expertise in mapping and dealing with ontology concepts.

Figure 15 shows the latest version of the model, a continuation of the development in protégé but with the participation of several experts as explained in section 4.4.3. This participation was phased by each CNS area, and taking the previous version as a starting point, improvements suggested by the different CNS clinical areas were discussed and implemented. This collaboration with the CNS resulted in a very significant evolution of the model which gained a greater depth (depth of 7) and a much wider conceptual scope. The active participation of physicians in numerical terms resulted in an increase of 86 concepts and 634 relationships, which represents an improvement in the order of 73% of concepts and 139% of relationships.

The final semantic model gathers 204 concepts, which is a model that balances specificity for a medical and biomedical context, with the wide coverage it has in this type of data, which allows that in the future the addition of new questionnaires does not require the creation of a new semantic model. This model was built together with the CNS doctors over a month and a half and therefore follows their mental model. The approval of this team of experts gives value and confidence to the model developed.



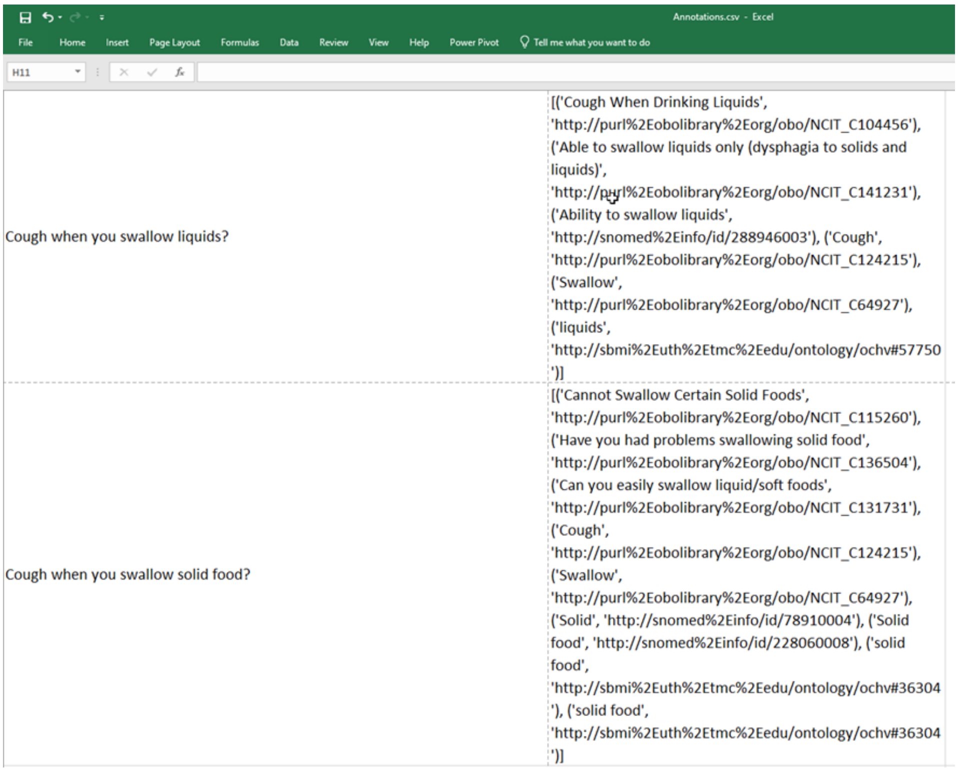
- Daily life
  - Autonomy
  - Daily activities
    - Dressing
    - Hygiene
    - Lying in bed
    - Reading
    - Sit
    - Sleep
    - Stand
    - Walking
    - Writing
  - Fatigue
  - Mood
    - Angriness
    - Anxiousness
    - Depressive
    - Sadness
    - Shame
    - Tiredness
  - Social Life
    - Oral Expression
      - Basic Expression
      - Complex Expression
      - Counting
      - Mental Recall (Evocation)
      - Naming
      - Repetition
      - Serial Speech
- Diagnostic Area
  - Neuropsychology
  - Nursing
  - Nutrition
  - Occupational Theraphy
  - Physiotherapy
  - Speech Therapy
- Disorder
  - Mental disorder
    - Anhedonia
    - Behavioural problems
    - Constructional Apraxia
    - Mood disorder
      - Anxiety
      - Apathy
      - Depression
    - Neurological disorders
      - Cognitive difficulties
        - Dementia
      - Dyskinesia
      - Hallucinations
      - Motor Symptoms
        - Dystonia
        - Fall
          - Anterior Fall
          - Posterior Fall
        - Fluctuations/ON/OFF
        - Gait
          - Festination
          - Freezing
          - slowness
        - Tremor
          - Kinetic Tremor
          - Postural Tremor
          - Resting Tremor
      - Non Motor Symptoms
        - Dizziness
        - Urinary System Disorder
          - Incomplete bladder emptying
          - Incontinence
          - Urinary Urgency
      - Psychosis
      - Sleep Problems
        - Insomnia
        - Obstructive Sleep Apnea
        - REM Sleep Behavior Disorder
        - Restless Legs Syndrome
        - Sleepiness
  - Physical disorder
    - Feeding difficulties
      - Dysphagia
        - Chewing difficulties
        - Swallowing disorder
    - Gastrointestinal disorders
      - Diarrhoea
      - Intestinal constipation
    - Language Disorder
      - Aphasia (Palilalia)
    - Motor Dysfunction
      - Akinesia
      - Bradykinesia
      - Dystonia
      - Orthostatic Hypotension
    - oral disfunction
      - Lip Disorder
      - Tongue Disorder
    - Respiratory disorder
      - Breathing difficulties
      - Bronchitis
      - Respiratory infections
    - Sarcopenia
      - Muscle depletion
      - Muscle weakness
      - Sarcopenic Obesity
    - Speech disorder
      - Cluttering (Tachyphemia)
      - Dysarthria
      - Dysphonia
      - Slurred speech
    - Visual Difficulty
      - Vision Dysfunction
    - Voice disorders
      - Hypophonia
- Medication
  - Taking Levodopa
- Mental state
  - Cognitive function
    - Abstraction
    - Attention
    - Calculation
    - Concentration
    - Language
      - Comprehension
      - Oral Expression
        - Basic Expression
          - Answer Questions
          - Yes or No Response
        - Complex Expression
          - Conversation
            - Group conversation
            - One-to-one conversation
            - Spontaneous speech
        - Counting
        - Mental Recall (Evocation)
        - Naming
        - Repetition
        - Serial Speech
    - Memory
      - Remote Memory
      - Verbal Memory
      - Visual Memory
      - Working Memory
    - Orientation
    - Retention
  - Emotion Regulation
  - Executive Function
  - Fatigue
  - Mental disorder
  - Mood
- Nutritional Status
  - Anthropometry
    - Body measures
      - Appendicular skeletal muscle mass
      - Basal metabolic rate
      - Body fat mass index
      - Dry lean weight
      - Extra-celular body water
      - Fat mass
      - Fat weight
      - Fat-free mass (index)
      - Healthy weight
      - Intra-celular body water
      - Lean mass
      - Lean weight
      - TBW(l)
      - Total body water
  - Diet
  - Hydration
    - Water Intake
  - Malnutrition
  - Nutritional intake
  - Nutritional Risk
  - Obesity
  - Pre-Obesity
- Physical state
  - Body measures
  - Fatigue
  - Frailty
  - Mobility
    - Balance
    - Dexterity
    - Effort Tolerance
    - Fall
      - Anterior Fall
      - Posterior Fall
    - Gait
      - Festination
      - Freezing
      - slowness
    - Involuntary Movements
      - Dyskinesia
    - Lifting
    - Postural Instability
    - Posture
      - Dystonia
    - Rigidity
    - Risk of Falling
    - Walking
  - Physical disorder
    - Feeding difficulties
      - Dysphagia
        - Chewing difficulties
        - Swallowing disorder
    - Gastrointestinal disorders
      - Diarrhoea
      - Intestinal constipation
    - Language Disorder
      - Aphasia (Palilalia)
    - Motor Dysfunction
      - Akinesia
      - Bradykinesia
      - Dystonia
      - Orthostatic Hypotension
    - oral disfunction
      - Lip Disorder
      - Tongue Disorder
    - Respiratory disorder
      - Breathing difficulties
      - Bronchitis
      - Respiratory infections
    - Sarcopenia
      - Muscle depletion
      - Muscle weakness
      - Sarcopenic Obesity
    - Speech disorder
      - Cluttering (Tachyphemia)
      - Dysarthria
      - Dysphonia
      - Slurred speech
    - Visual Difficulty
      - Vision Dysfunction
    - Voice disorders
      - Hypophonia
  - Physical functions
    - Communication
      - Speech
        - Intelligibility
        - Voice
          - Laryngeal Function
    - Deglutition
    - Elimination
      - Gastrointestinal function
      - Urinary system
    - Sexual Function
  - Physical responses
    - Appetite
    - Nausea
    - Pain
    - Shortness of breath
    - Sialorrhoea
    - Syncope
    - Xerostomia
  - Reactance
  - Resistance
  - Saliva Health
    - Sialorrhoea
    - Xerostomia
  - Skin Integrity
  - Strength
  - Vital Signs
    - Arterial Pressure
    - Body Temperature
    - Pulse Rate
    - Respiratory Rate
- Transfers
  - Lying in bed
  - Sit
  - Stand

Figure 15: Final result of several versions of the model evolution.

5.4 Semantic Annotation

As far as annotation is regarded, as seen in section 4.6 Semantic Annotation of Questionnaires, two annotators, BioPortal Annotator and ElasticSearch were used.

The performance of the automatic annotation was evaluated using a method similar to the machine translation evaluation. As such, was proceeded with the automatic annotation of a built corpus of sixty questions, the output of the program is a CSV file, which associates to each question in the corpus a list of tuples containing a pair consisting of the ontology concept and its URI.



Cough when you swallow liquids?	[('Cough When Drinking Liquids', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C104456'), ('Able to swallow liquids only (dysphagia to solids and liquids)', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C141231'), ('Ability to swallow liquids', 'http://snomed%2Einfo/id/288946003'), ('Cough', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C124215'), ('Swallow', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C64927'), ('liquids', 'http://sbmi%2Euth%2Etmc%2Eedu/ontology/ochv#57750')]
Cough when you swallow solid food?	[('Cannot Swallow Certain Solid Foods', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C115260'), ('Have you had problems swallowing solid food', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C136504'), ('Can you easily swallow liquid/soft foods', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C131731'), ('Cough', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C124215'), ('Swallow', 'http://purl%2Eobolibrary%2Eorg/obo/NCIT_C64927'), ('Solid', 'http://snomed%2Einfo/id/78910004'), ('Solid food', 'http://snomed%2Einfo/id/228060008'), ('solid food', 'http://sbmi%2Euth%2Etmc%2Eedu/ontology/ochv#36304'), ('solid food', 'http://sbmi%2Euth%2Etmc%2Eedu/ontology/ochv#36304')]

Figure 16: CSV output of questions semantic annotation generated by the program.

Figure 16 shows as an example two questions, and their concept annotations, for better understanding let's dissect the question "Cough when you swallow solid food?".

Explaining the annotation of the question "Cough when you swallow solid food?" (Figure 17), when mapping this question through the program, an output is generated consisting of a list of nine tuples of terms with their URI. Each of the terms has an associated link(URI) that refers to the ontological details of the concept in one of the four ontologies selected in Section 5.2 Ontology Set.

Since this is an automatic annotation, some flaws are to be expected, and in this specific case of the nine annotated concepts one is considered invalid - "Solid"- since it does not refer to the intended medical/biomedical context nor to any scenario that helps semantically enrich the question under analysis.

Then there are three repeated concepts - "Solid food" - of which only one has semantic value, that of the snomed ontology. The remaining concepts are all valid given the context of the question and also have a semantic value that allows to relate them to other medical/biomedical concepts that will allow bridging other questions in other questionnaires or even in other diagnostic areas.

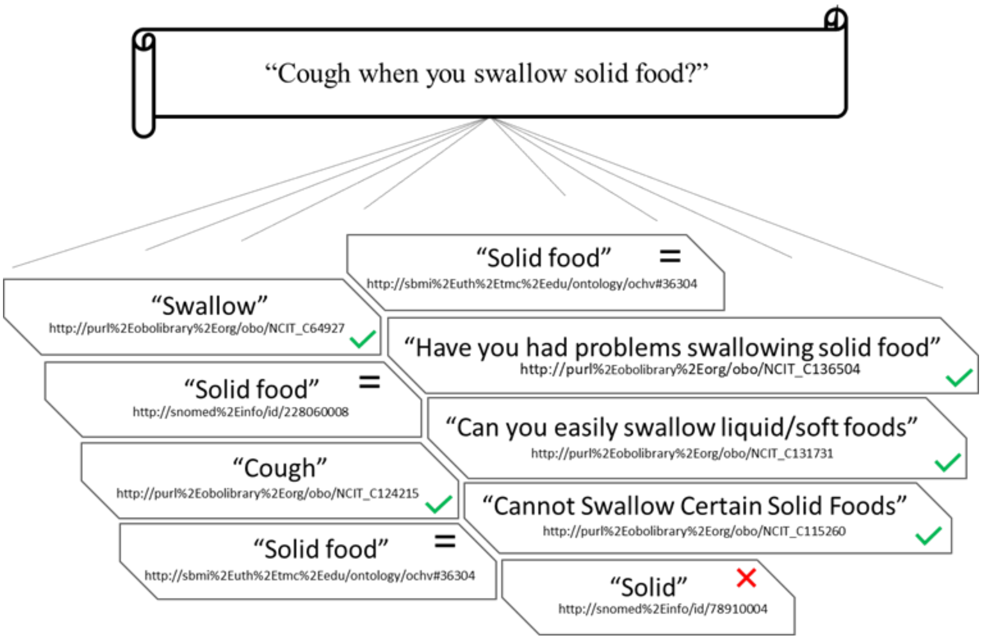

“Cough when you swallow solid food?”
“Solid food” =
http://sbmi%2Euth%2Etmc%2Eedu/ontology/ochv#36304
“Swallow”
http://purl%2Eobolibrary%2Eorg/obo/NCIT_C64927 ✓
“Have you had problems swallowing solid food”
http://purl%2Eobolibrary%2Eorg/obo/NCIT_C136504 ✓
“Solid food” =
http://snomed%2Einfo/id/228060008
“Can you easily swallow liquid/soft foods”
http://purl%2Eobolibrary%2Eorg/obo/NCIT_C131731 ✓
“Cough”
http://purl%2Eobolibrary%2Eorg/obo/NCIT_C124215 ✓
“Cannot Swallow Certain Solid Foods”
http://purl%2Eobolibrary%2Eorg/obo/NCIT_C115260 ✓
“Solid food” =
http://sbmi%2Euth%2Etmc%2Eedu/ontology/ochv#36304
“Solid” ✗
http://snomed%2Einfo/id/78910004

Figure 17: Annotation output informal example.

In summary, the automatic annotation resulted in nine mapped terms, of which eight are valid terms given the context of the question, seven have semantic value, and there are three repeated terms, as two of these 3 concepts are exactly equal (same term and same URI) only one of them are considered.

After this survey was done about the output of the program, the manual annotation of the question was done, where three concepts that allow the total mapping of the question were obtained, and only with semantic value concepts. Therefore, the minimum number of automatic annotations must not be less than the number of manual annotations, for the annotation to be considered correct.

Further, from the nine mapped concepts, there is a hit of six in ten, and it should be noted that the question was fully annotated with no missing terms, manually only three terms have allowed the complete annotation of this question. As such, by doing the percentage of valid concepts with and without repetition and also the percentage of terms with the semantic value it can get a more concrete idea about the performance of the built pipeline.

To get a more robust evaluation, the process done for this question was replicated for the entire corpus of 60 questions, not only all the outputs of the pipeline were evaluated, but in parallel, all the questions were manually annotated and subsequently compared and evaluated as successful annotation or not. Automatic annotation was considered successful if its output has several valid terms without duplicates equal to or greater than that obtained manually (valid terms are terms that make sense in the question context).

The behaviour exemplified above for a question is replicated in the entire corpus of 60 questions. Looking in detail at the results of the automatic annotations performed by the developed pipeline I chose to analyse the behaviour of the three previously mentioned parameters that give value to the annotations: i) The number of annotated terms considered valid, ii) The number of terms considered valid, but without duplicates and iii) The number of annotated terms with semantic value.

In summary, a term is considered valid when it fits the context of the question in question, i.e. the mapping that is made between the ontological term and the term present in the question makes sense and corresponds to the reality of the question. The occurrence of invalid terms is higher in homonymous words.

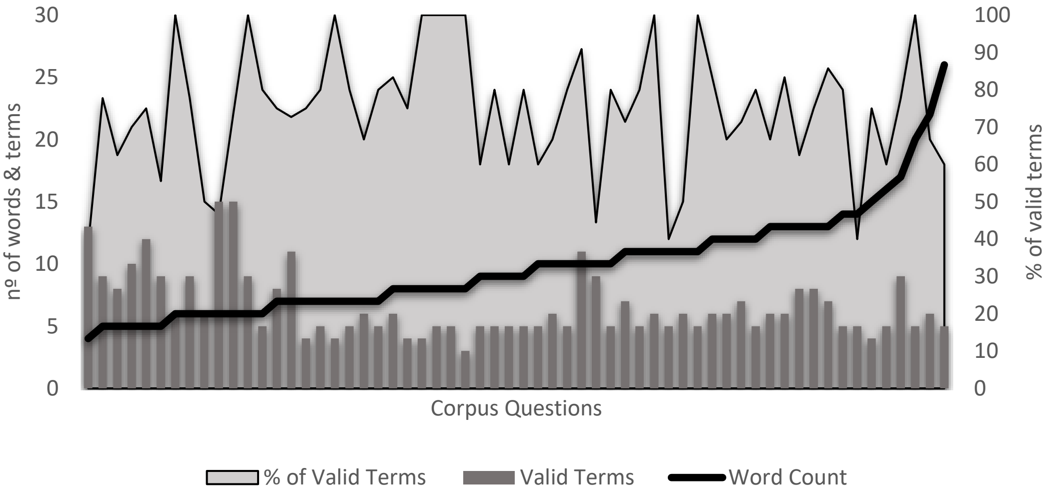


Figure 18: Variation of number (absolute and percentage) of terms with the variation of the number of words of each question in the corpus.

Figure 18 relates the number of words in the question with the number and percentage of valid terms in each question. It's visible that with the increase in the number of words in a question there is a tendency for the % of valid terms per question to decrease and that the number of valid terms does not accompany the number of words in the question. Questions with between 4 and 10 words had more success in the annotation of valid terms and also in the % of annotated terms that are considered valid in relation to the number of words in the question. While the average % of valid terms is 75.6% in questions with a number of words between 4 and 10, questions with more than 10 words have a lower percentage of 72.8%, the average of annotated terms considered valid is in the same order of 7.1 valid terms vs 5.9 in questions with more than 10 words.

Figure 18 also shows that in questions with fewer words, several annotated terms exceed the line of the number of words that the question has, this happens because as the example of Figure 17 shows, for the same word of the question can have the annotation repeated or have different annotations that refer to the same ontological term but from different ontologies. Considering this factor, I decided to investigate according to the previous logic but now removing the repeated terms, so in Figure 19 we already see the results without repeated terms.

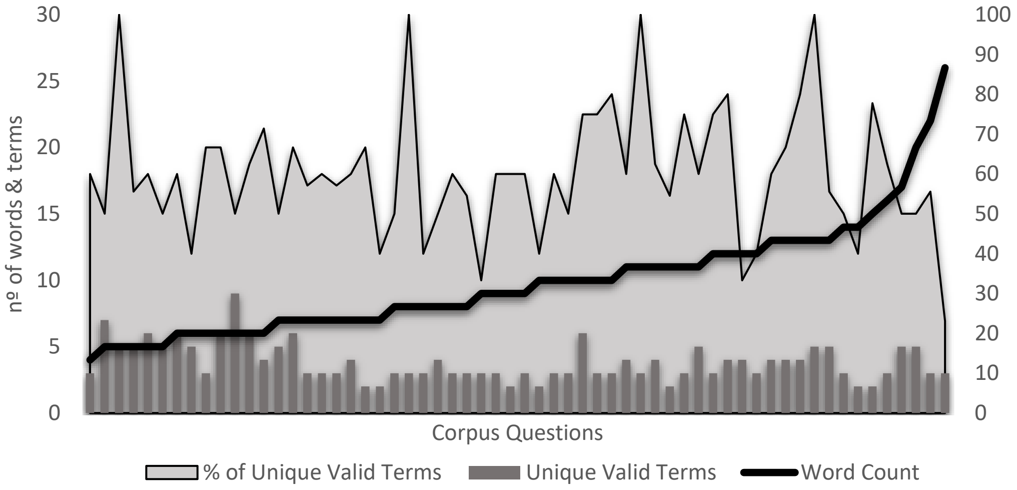


Figure 19: Variation of the number (absolute and percentage) of unique valid terms with the variation of the number of words of each question in the corpus

Figure 19 have the results of the relationship between the number of words in the question and the number and percentage of valid unique terms for each question. A term is considered unique if the output generated for that question presents the URI only once. In all cases in which the URI appears repeated, the duplicates are removed, leaving only a single term referring to a given URI.

With this data filter, there was naturally a marked fall both in the percentage of single valid terms per question and in the average number of valid terms (now without repeats); for questions with up to 10 words there was a 15.9% decrease (59.7% of the terms annotated are single valid terms), and the average number of terms fell from 7.1 to 4.0. In questions consisting of more than 10 words, the decrease is slightly lower, 12.0%

(60.9% of the terms noted are valid single terms), and the average number of terms fell from 5.9 to 3.6.

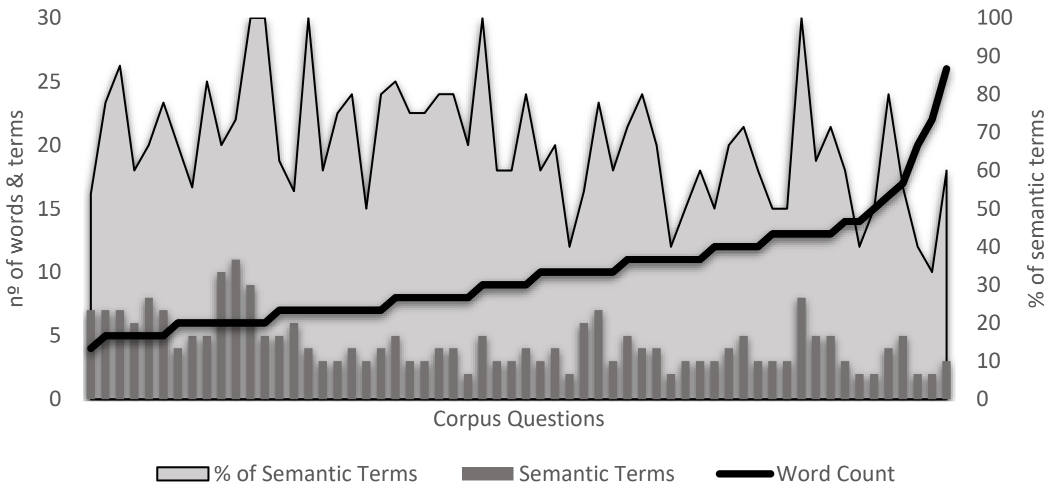


Figure 20: Variation of the number (absolute and percentage) of semantic value terms with the variation of the number of words of each question in the corpus.

About the identification of annotated terms with semantic value, the graphic presented in Figure 20 shows that at the percentage level there is a higher percentage of this type of terms in questions with 10 or fewer words, 71.6% against 59.5% in questions with more than 10 words. At the level of the average of terms with a semantic value per question, it has values of 5.0 for questions of shorter length and 3.6 for longer questions. Once again, the number of words does not seem to influence the number of annotations with semantic value.

The 3 figures above show that the pipeline performs better for shorter questions and that the possibility of dispersion of the annotation increases with the number of words in the question. Longer questions also have potentially more stop words, which can hinder correct annotation in a specific context. Smaller questions have fewer stop words and naturally a higher % of relevant constituent words.

Finally, it was made a survey of the questions with successful or unsuccessful annotation. In 8 cases I classified the question as badly annotated since the pipeline annotated the question with several valid terms lower than with the manual annotation. Having the corpus of 60 questions, the percentage of questions well annotated by the program developed was 87%.

Figure 21 shows the comparison of the 52 well-annotated questions versus the 8 poorly annotated ones, to try to understand if there is any factor that contributes to the annotation failure. What immediately stands out is the fact that two of the three parameters under study have a significant decrease in their averages. The parameter that shows a greater difference is the unique valid terms annotated, which shows a difference of about 20%, followed by the valid terms parameter, which shows a decrease of about 15%. This may indicate that the identification of valid terms with and without repetitions has great importance in the success of the annotation. The other parameter of the identification of semantic terms remains practically constant, leading to the belief that it does not influence the quality of the annotation itself.

In summary, the program obtained a score of 87% for annotation success, with a higher performance on shorter questions. The most important factors identified for the success of the annotations are the number of valid terms and unique valid terms.

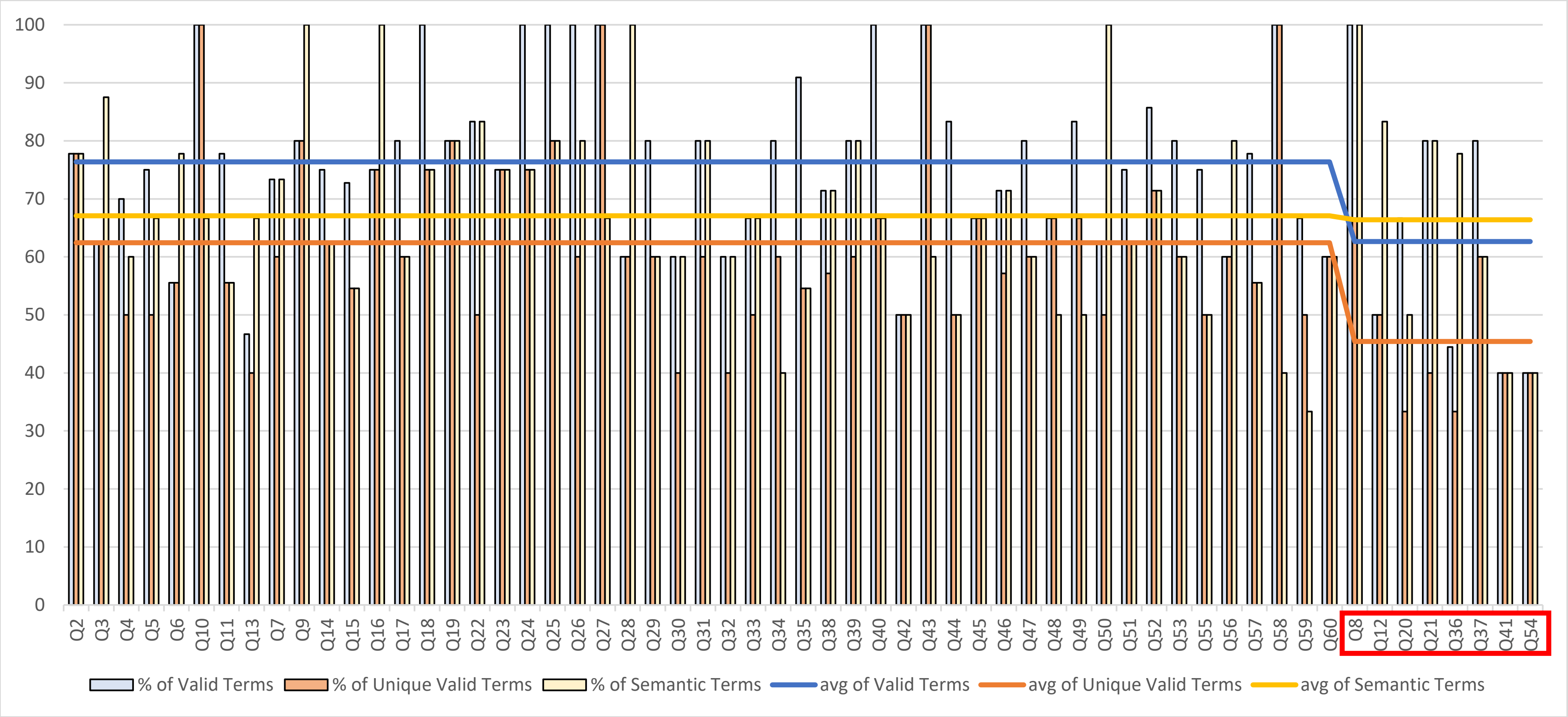


Figure 21: Influence of the parameters understudy on the quality of the annotations. (Questions inside the red rectangle are classified as unsuccessful annotations, all others are successful)

5.5 Semantic Integration

This diagram in Figure 22 represents the output generated by the built pipeline and shows how different questions are integrated through previous annotations and also the integration of patient data with these same questions. Therefore, after each question is annotated with the respective ontological concepts, it becomes possible to have an overview of all the questions in the different questionnaires at the concept level, i.e., if we only want to see answers that refer to a given concept, we can filter them to do so.

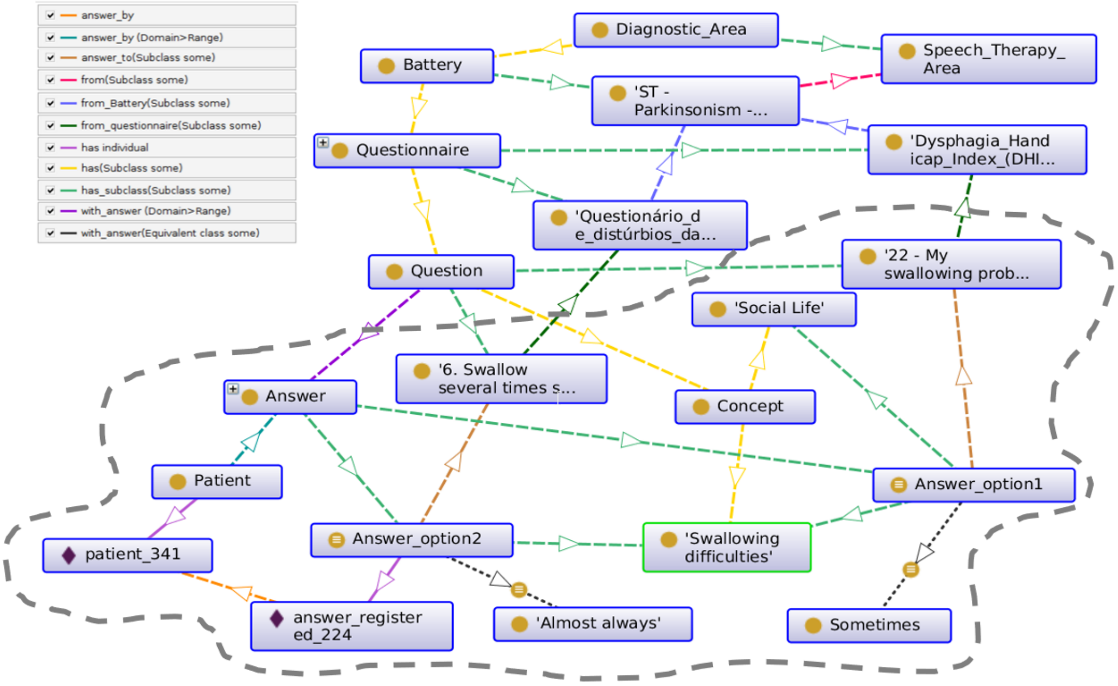


Figure 22: Informal demonstration of the overall functioning of the pipeline.

The potential of this possibility gains strength with the integration in the pipeline of patients' clinical data. So, a certain patient in the last one has a given number of consultations performed, during and between which he was filling out questionnaires that the doctor in charge thought pertinent. If before, to have a holistic view of the patient, the doctor had to go through several questionnaires, several questions and their answers, as well as their evolution over time, with a built pipeline this is no longer necessary. Now the doctor in charge can make a diagnosis, or assess the evolution of a given patient about a given concept. For example, the concept "Swallowing difficulties" that is present in several questions in different questionnaires has an easier time being tracked since all the questions that address this concept are marked with the same id (the Uri meaning " Swallowing difficulties") in the ontologies. Therefore, it can see all the questions that address this concept, as well as the answers that the patient gave to them, and this with the temporal evolution that allows tracing the tendency to improve or worsen about a given concept.

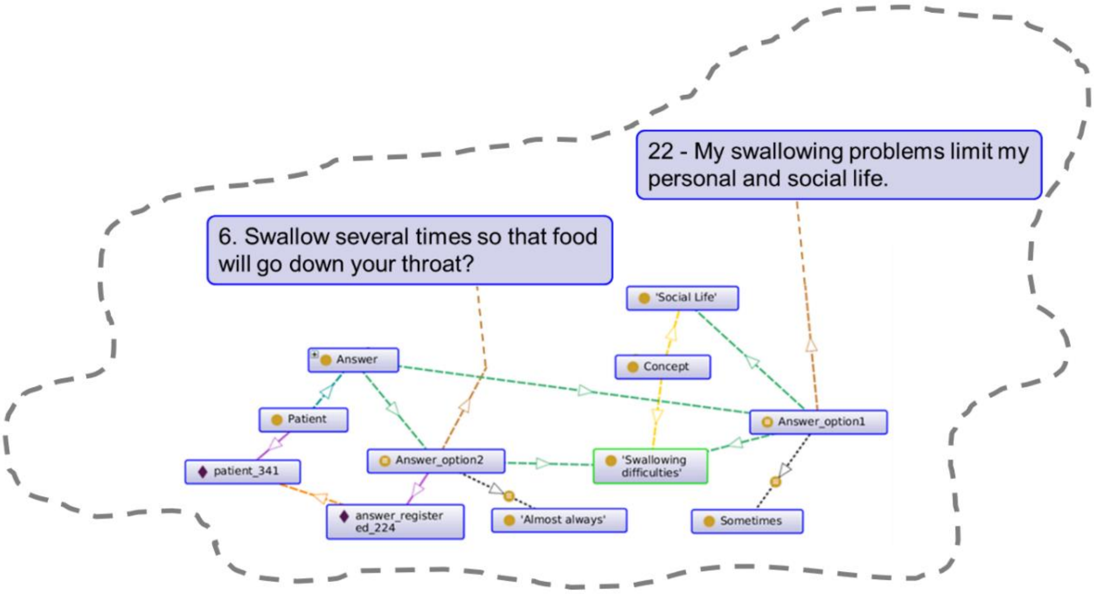
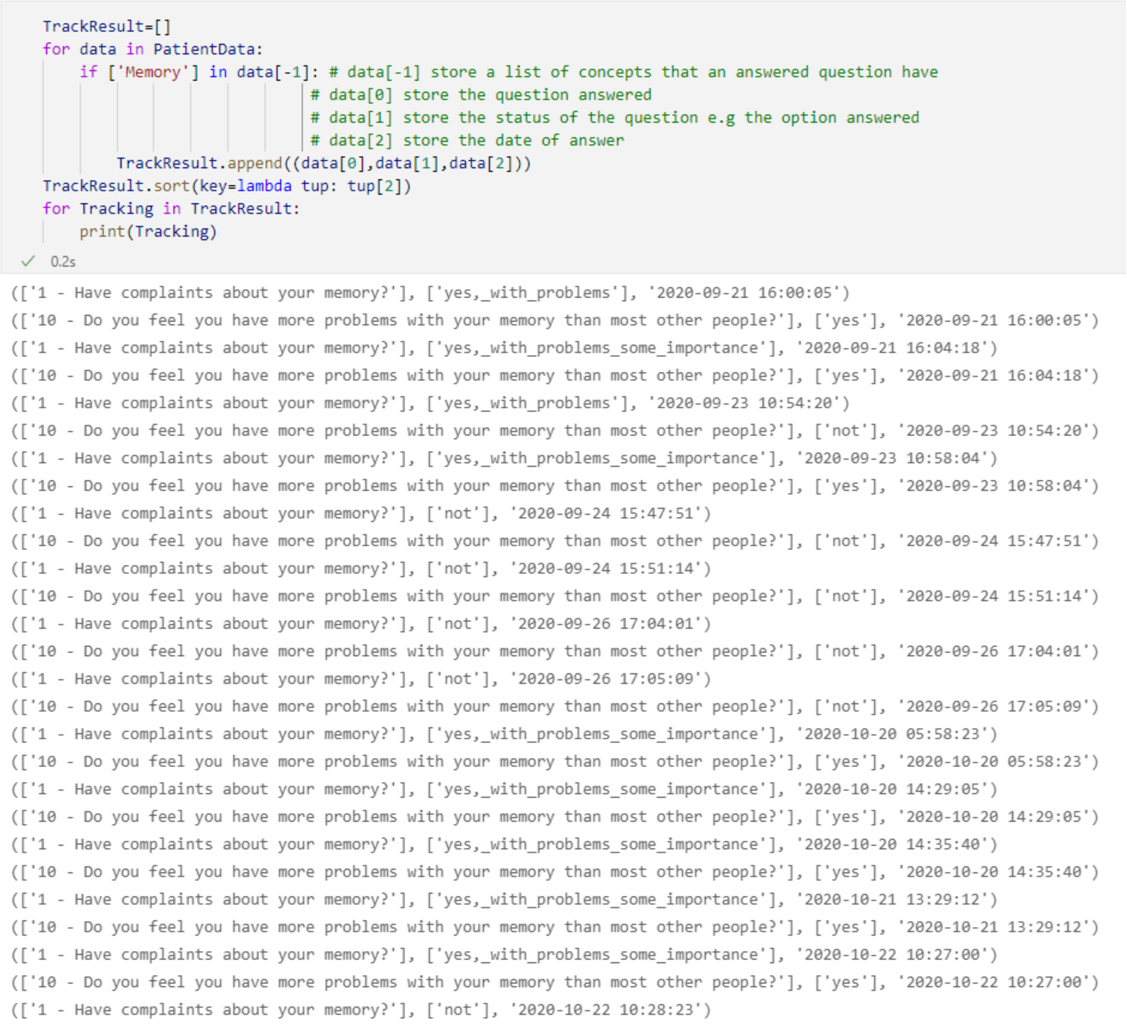

22 - My swallowing problems limit my personal and social life.
6. Swallow several times so that food will go down your throat?
'Social Life'
Answer
Concept
Patient
Answer_option1
patient_341
Answer_option2
'Swallowing difficulties'
answer_register
ed_224
'Almost always'
Sometimes

Figure 23: Detailed Informal demonstration of how the pipeline works about the annotation and semantic integration of two distinct questions and the cross-referencing with a given patient's data.

In this specific example, shown in detail in Figure 23 we have the integration of two different questions, from different questionnaires, i) "Swallow several times so that food will go down your throat?" and ii) "My swallowing problems limit my personal and social life.", in i) the program annotated the question with the term "swallowing difficulties" only, in ii) the program annotated the question with 2 terms "swallowing difficulties" and "social life". Thus, we have an intersection between these two questions, about the concept of "swallowing difficulties". This is just an illustrative example of how the developed program works, in this case, we have only two questions but in reality, the expected is that all questions that have been annotated with the concept "swallowing difficulties" point to this ontological term. Having said that, we then introduce the integration of data from a patient, in this case, patient 341, who has a recorded answer on a given date, that recorded answer is composed of the concepts annotated in the question concerned together with the given answer option. In this specific case the answer to the question was "Almost always" so this registered answer keeps the concept "swallowing difficulties" and the status "Almost always" on a given date, obviously this concept-status pair may vary over time, which may help the physician in the decision-making process in certain contexts.

The simulation of the holistic vision that the doctor could obtain with the work carried out in this dissertation was attempted. Considering the integration of the semantic model, with the 4 ontologies and the DataPark Data, where all the questions and respective annotations are inserted plus the patients' data, the OWL file that contains all this integrated information has an enormous weight close of 1.3 Gb.

To handle the large requirements of the model while still demonstrating the holistic integration abilities, a small proof-of-concept application was built in python to interrogate the semantic data and Figure 24 shows an example of the potential of these queries, which in this case allow tracking over time the answers to two questions that assess the concept of memory. Part of the future work will be to deploy the data in a triplestore and integrate the queries and exploration with user interfaces.

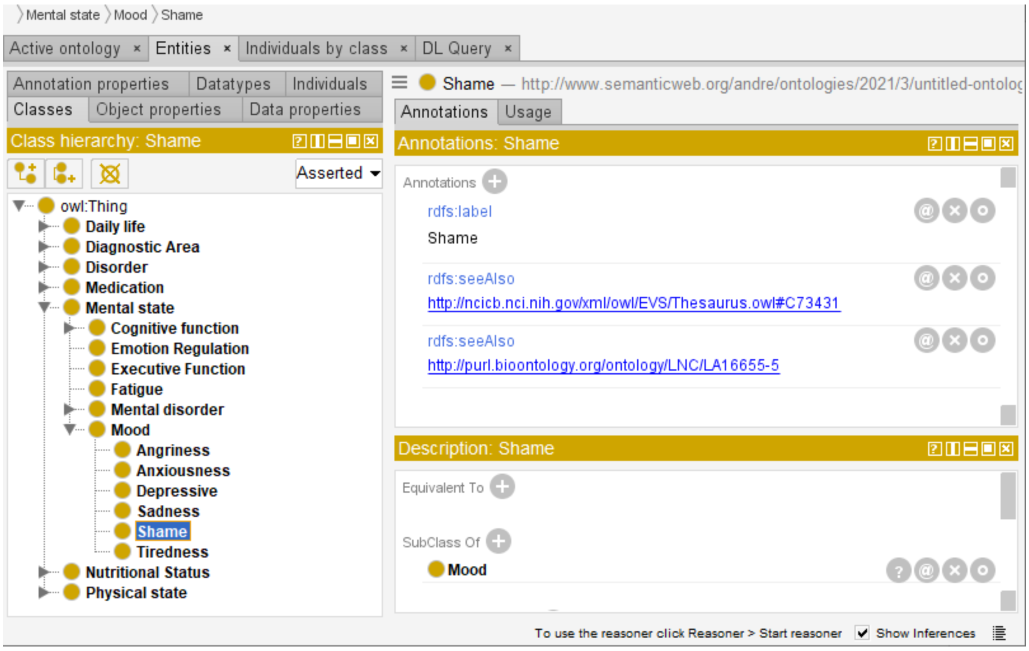


```
TrackResult=[]
for data in PatientData:
    if ['Memory'] in data[-1]: # data[-1] store a list of concepts that an answered question have
                               # data[0] store the question answered
                               # data[1] store the status of the question e.g the option answered
                               # data[2] store the date of answer
        TrackResult.append((data[0],data[1],data[2]))
TrackResult.sort(key=lambda tup: tup[2])
for Tracking in TrackResult:
    print(Tracking)
```

✓ 0.2s

```
(['1 - Have complaints about your memory?'], ['yes,_with_problems'], '2020-09-21 16:00:05')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-09-21 16:00:05')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-09-21 16:04:18')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-09-21 16:04:18')
(['1 - Have complaints about your memory?'], ['yes,_with_problems'], '2020-09-23 10:54:20')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['not'], '2020-09-23 10:54:20')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-09-23 10:58:04')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-09-23 10:58:04')
(['1 - Have complaints about your memory?'], ['not'], '2020-09-24 15:47:51')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['not'], '2020-09-24 15:47:51')
(['1 - Have complaints about your memory?'], ['not'], '2020-09-24 15:51:14')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['not'], '2020-09-24 15:51:14')
(['1 - Have complaints about your memory?'], ['not'], '2020-09-26 17:04:01')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['not'], '2020-09-26 17:04:01')
(['1 - Have complaints about your memory?'], ['not'], '2020-09-26 17:05:09')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['not'], '2020-09-26 17:05:09')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-10-20 05:58:23')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-10-20 05:58:23')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-10-20 14:29:05')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-10-20 14:29:05')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-10-20 14:35:40')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-10-20 14:35:40')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-10-21 13:29:12')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-10-21 13:29:12')
(['1 - Have complaints about your memory?'], ['yes,_with_problems_some_importance'], '2020-10-22 10:27:00')
(['10 - Do you feel you have more problems with your memory than most other people?'], ['yes'], '2020-10-22 10:27:00')
(['1 - Have complaints about your memory?'], ['not'], '2020-10-22 10:28:23')
```

Figure 24: python queries over patient data.



〉Mental state 〉Mood 〉Shame

Active ontology × | Entities × | Individuals by class × | DL Query ×

Annotation properties | Datatypes | Individuals
Classes | Object properties | Data properties

Class hierarchy: Shame

Asserted

- owl:Thing
  - Daily life
  - Diagnostic Area
  - Disorder
  - Medication
  - Mental state
    - Cognitive function
    - Emotion Regulation
    - Executive Function
    - Fatigue
    - Mental disorder
    - Mood
      - Angriness
      - Anxiousness
      - Depressive
      - Sadness
      - Shame
      - Tiredness
  - Nutritional Status
  - Physical state

Shame — http://www.semanticweb.org/andre/ontologies/2021/3/untitled-ontolog

Annotations | Usage

Annotations: Shame

Annotations

rdfs:label
Shame

rdfs:seeAlso
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C73431

rdfs:seeAlso
http://purl.bioontology.org/ontology/LNC/LA16655-5

Description: Shame

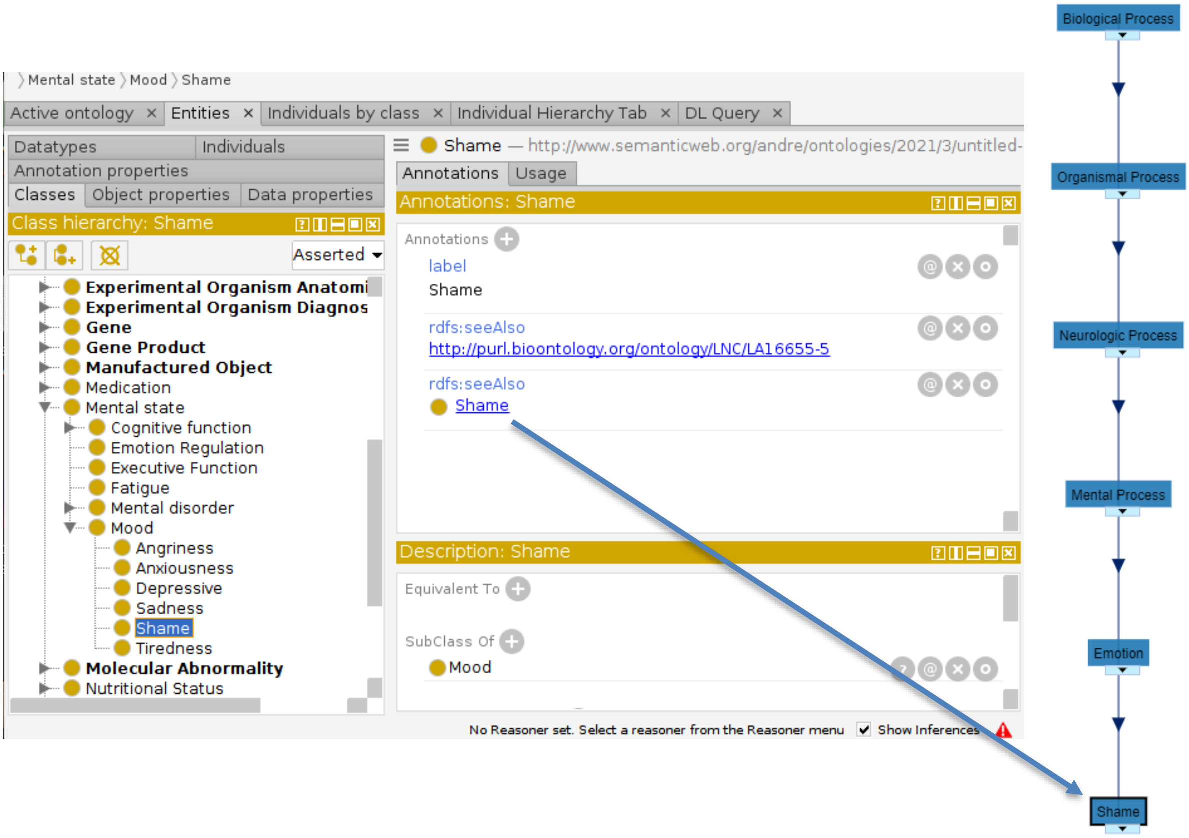
Equivalent To

SubClass Of
Mood

To use the reasoner click Reasoner > Start reasoner ☑ Show Inferences

Figure 25: Semantic Model without ontology integration.

What the designed pipeline allows is a full integration of the semantic model, the selected ontologies and the DataPark data with the respective annotations. Figure 25 shows the semantic model without the integration with the ontologies and Figure 26 with the integration with the NCIT ontology, which allows you to see was the hierarchy of



〉Mental state 〉Mood 〉Shame

Active ontology × | Entities × | Individuals by class × | Individual Hierarchy Tab × | DL Query ×

Datatypes | Individuals
Annotation properties
Classes | Object properties | Data properties

Class hierarchy: Shame

Asserted

- Experimental Organism Anatomi
- Experimental Organism Diagnos
- Gene
- Gene Product
- Manufactured Object
- Medication
- Mental state
  - Cognitive function
  - Emotion Regulation
  - Executive Function
  - Fatigue
  - Mental disorder
  - Mood
    - Angriness
    - Anxiousness
    - Depressive
    - Sadness
    - Shame
    - Tiredness
- Molecular Abnormality
- Nutritional Status

Shame — http://www.semanticweb.org/andre/ontologies/2021/3/untitled-

Annotations | Usage

Annotations: Shame

Annotations

label
Shame

rdfs:seeAlso
http://purl.bioontology.org/ontology/LNC/LA16655-5

rdfs:seeAlso
Shame

Description: Shame

Equivalent To

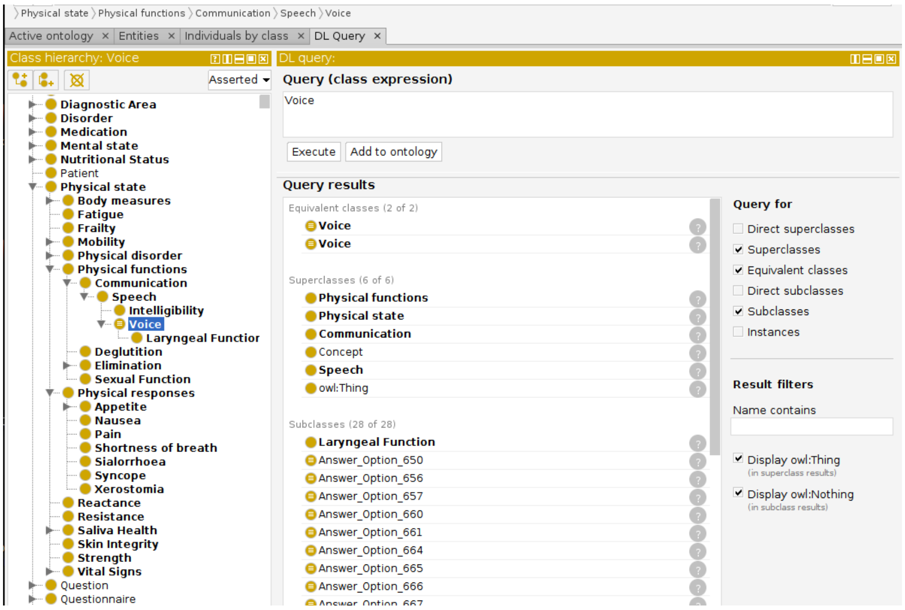
SubClass Of
Mood

No Reasoner set. Select a reasoner from the Reasoner menu ☑ Show Inferences

Figure 26: Semantic Model with NCIT integration.

“shame” in the NCIT, Shame<Emotion<Mental Process<Neurological process. These 2 Figures show that the merge between the model and the NCIT ontology occurs when the URI are equal (URI is replaced by ontology term label), i.e. the model term was annotated with the term present in the NCIT, thus the link gives.

What this link to external ontologies allows is all the knowledge networks contained therein, superclasses, subclasses etc.



```
▶ Diagnostic Area
▶ Disorder
▶ Medication
▶ Mental state
▶ Nutritional Status
  Patient
▼ Physical state
    ▶ Body measures
      Fatigue
      Frailty
    ▶ Mobility
    ▶ Physical disorder
    ▼ Physical functions
        ▼ Communication
            ▼ Speech
                  Intelligibility
                ▼ Voice
                      Laryngeal Functior
          Deglutition
        ▶ Elimination
          Sexual Function
    ▼ Physical responses
        ▶ Appetite
          Nausea
          Pain
          Shortness of breath
          Sialorrhoea
          Syncope
          Xerostomia
      Reactance
      Resistance
    ▶ Saliva Health
      Skin Integrity
      Strength
    ▶ Vital Signs
▶ Question
▶ Questionnaire
```

Query (class expression)

```
Voice
```

Query results

```
Equivalent classes (2 of 2)
    Voice
    Voice

Superclasses (6 of 6)
    Physical functions
    Physical state
    Communication
    Concept
    Speech
    owl:Thing

Subclasses (28 of 28)
    Laryngeal Function
    Answer_Option_650
    Answer_Option_656
    Answer_Option_657
    Answer_Option_660
    Answer_Option_661
    Answer_Option_664
    Answer_Option_665
    Answer_Option_666
    Answer_Option_667
```

Query for

☐ Direct superclasses
☑ Superclasses
☑ Equivalent classes
☐ Direct subclasses
☑ Subclasses
☐ Instances

Result filters

Name contains

☑ Display owl:Thing (in superclass results)
☑ Display owl:Nothing (in subclass results)

Figure 27: SemanticModel + NCIT + DataPark integrated.

In this last Figure 27, it presents the result of a very simple query on the semantic model integrated with the NCIT(Laryngeal Function is NCIT subclass) plus the DataPark Data, in this case, it was searched for the term "Voice" and can see equivalents class, all superclasses and subclasses present in the 3 elements of integration (ontology + model + DataPark), so also appear all the response options that have the annotation to the question with the concept "Voice".

Chapter 6

Conclusion

In Chapter 1 (Introduction) I presented the DataPark, a platform that had the potential for a new vision by physicians. This implied a changing paradigm with which CNS physicians used the DataPark questionnaires and scales. Before this work, DataPark vision was specific by questionnaire and area of diagnosis, i.e. doctors could only follow the evolution of the patient about a given questionnaire. The proposed was to change this view of doctors across the questionnaires, changing the view at the level of questionnaires and give to Doctors a holistic view at the level of the concept. The developed program allowed this huge change because it allows them to use all the data from the questions, whether they are from questionnaires or different diagnostic areas or not. In this way, Doctors have a concept-oriented view of the patient.

All the processes to reach this proposed objective derived from challenges that crossed the study path to the final objective, among which was the automatic translation of all the questionnaires, a crucial point of this dissertation, since only a quality translation can produce correct annotations and, consequently, a good semantic integration of the data. To be able to evaluate the translation process, I had to create a Gold Standard Corpus from which I could measure the performance of translation, which obtained a precision of 0.78, a recall of 0.76 and an F1 of 0.77, results that give me the confidence to proceed to the semantic annotation. Here, to evaluate the performance, I resorted to the manual annotation that served as a reference to compare with the results generated by the program, also here I obtained a safe result of 87% of the annotations being classified as successful.

The development of the semantic model was also a success, initially developed with the help of LASIGE and DataPark researchers and, in the most important phase, with the participation of 9 CNS experts among physicians and therapists, who supported a constant revision and improvement of the model during 9 zoom meetings. The developed model reached a total of 204, with 1089 relations established and 436 links to external ontologies. The participation of the CNS clinical team provided greater robustness to this model.

Once the semantic model, the processes of translation and annotation with ontologies were concluded, it was then possible to integrate all the questions of the different questionnaires, this integration is done at the concept level. A given concept then refers to a certain number of questions, and after integration with the patients' real data, it is possible to relate all the patients' statuses regarding the most varied concepts and their evolution over time.

Therefore, all the objectives proposed for this dissertation were achieved and a base was built that will allow, in the future, for doctors to have support for decision making and for the analysis of the patient's evolution throughout time.

6.1 Future Work

Some future work can focus on some of the limitations of the current approach. There is room to improve the results of the semantic annotation using deep learning-based techniques, such as BioBert which were not used in the proposed methodology. The initial goal was to build a pipeline made of openly available and easy to integrate semantic tools. Although the limitations of the NCBO BioPortal Annotator prompted the development of a straightforward annotator based on ElasticSearch, this approach is still word and string-based, and therefore can fail to properly annotate input text, especially in regards to homonymy and synonymy. The model itself will require updates to increase the coverage of other questionnaires and scales not considered in the dataset used. However, the most relevant next step is the development of user-friendly visualizations and reports that allow healthcare providers to have a holistic view of the patient, which instead of being organized by domain, test battery and questionnaire, is concept-oriented. Before this work, the data in DataPark was only accessible by navigating between different test batteries and questionnaires. Now, data can be browsed and queried by concept using semantic annotations, which opens up new avenues to improve communication across expert teams and support coordination efforts.

References

Amith, M. *et al.* (2018) “Assessing the practice of biomedical ontology evaluation: Gaps and opportunities,” *Journal of Biomedical Informatics*. Academic Press Inc., pp. 1–13. doi:10.1016/j.jbi.2018.02.010.

Aroyo, L. *et al.* (2010) “Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6089 LNCS(PART 2). doi:10.1007/978-3-642-13489-0.

Bai, L. *et al.* (2021) “Clinical Entity Extraction: Comparison between MetaMap, cTAKES, CLAMP and Amazon Comprehend Medical,” in *2021 32nd Irish Signals and Systems Conference, ISSC 2021*. Institute of Electrical and Electronics Engineers Inc. doi:10.1109/ISSC52156.2021.9467856.

Beasley, L. and Manda, P. (2018) “Comparison of natural language processing tools for automatic Gene Ontology annotation of scientific literature,” *CEUR Workshop Proceedings*, 2285, pp. 1–7. doi:10.7287/peerj.preprints.27028.

Browne, A.C. *et al.* (2003) “UMLS Language and Vocabulary Tools AMIA 2003 Open Source Expo,” *AMIA 2003 Symposium Proceesings*, (July 2014), p. 798.

Cheatham, M. and Pesquita, C. (2017) “Semantic data integration,” *Handbook of Big Data Technologies*, pp. 263–305. doi:10.1007/978-3-319-49340-4_8.

Devlin, J. *et al.* (no date) *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Available at: https://github.com/tensorflow/tensor2tensor.

Dhayne, H. *et al.* (2019) “In Search of Big Medical Data Integration Solutions - A Comprehensive Survey,” *IEEE Access*, 7, pp. 91265–91290. doi:10.1109/ACCESS.2019.2927491.

Diogo Branco, César Mendes, Ricardo Pereira, André Rodrigues, Raquel Bouça-Machado, Kyle Montague, Joaquim Ferreira, T.G. (2019) “DataPark: A Data-Driven Platform for Parkinson’s Disease Monitoring,” *WISH Symposium - Workgroup on Interactive Systems in Healthcare, co-located with CHI’19* [Preprint], (April). Available at: https://www.researchgate.net/publication/332471204_DataPark_A_Data-Driven_Platform_for_Parkinson’s_Disease_Monitoring%0Ahttps://tjvguerreiro.github.io/pubs/ddhp_wish.pdf.

Dugas, M. *et al.* (2016) “ODMedit: Uniform semantic annotation for data integration in medicine based on a public metadata repository,” *BMC Medical Research Methodology*, 16(1), pp. 1–9. doi:10.1186/s12874-016-0164-9.

Funk, C. *et al.* (2014) “Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters,” *BMC Bioinformatics*, 15(1). doi:10.1186/1471-2105-15-59.

Galeota, E. and Pelizzola, M. (2017) “Ontology-based annotations and semantic relations in large-scale (epi)genomics data,” *Briefings in Bioinformatics*, 18(3), pp. 403–412. doi:10.1093/bib/bbw036.

Garcia, A. *et al.* (2018) “Biotea: Semantics for Pubmed Central,” *PeerJ*, 2018(1), pp. 1–26. doi:10.7717/peerj.4201.

Gruber, T.R. (1995) “Toward principles for the design of ontologies used for knowledge sharing,” *International Journal of Human - Computer Studies*, 43(5–6), pp. 907–928. doi:10.1006/ijhc.1995.1081.

Haendel, M.A., Chute, C.G. and Robinson, P.N. (2018) “Classification, Ontology, and Precision Medicine,” *New England Journal of Medicine*, 379(15), pp. 1452–1462. doi:10.1056/nejmra1615014.

Hoehndorf, R., Schofield, P.N. and Gkoutos, G. v. (2015) “The role of ontologies in biological and biomedical research: A functional perspective,” *Briefings in Bioinformatics*, 16(6), pp. 1069–1080. doi:10.1093/bib/bbv011.

Jayaratne, M. *et al.* (2019) “A data integration platform for patient-centered e-healthcare and clinical decision support,” *Future Generation Computer Systems*, 92(September), pp. 996–1008. doi:10.1016/j.future.2018.07.061.

Jonquet, C., Shah, N.H. and Musen, M.A. (2009) “The open biomedical annotator.,” *Summit on translational bioinformatics*, 2009, pp. 56–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21347171%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3041576.

Jovanović, J. and Bagheri, E. (2017) “Semantic annotation in biomedicine: The current landscape,” *Journal of Biomedical Semantics*, 8(1), pp. 1–18. doi:10.1186/s13326-017-0153-x.

Kersloot, M.G. *et al.* (2020) “Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies,” *Journal of Biomedical Semantics*. BioMed Central Ltd. doi:10.1186/s13326-020-00231-z.

Larmande, P. and Jibril, K.M. (2020) “Enabling a fast annotation process with the table2annotation tool,” *Genomics and Informatics*, 18(2), pp. 42–47. doi:10.5808/GI.2020.18.2.e19.

Lee, J. *et al.* (2020) “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, 36(4), pp. 1234–1240. doi:10.1093/bioinformatics/btz682.

Liao, Y. *et al.* (2011) “Why , Where and How to use Semantic Annotation for Systems Why , Where and How to use Semantic Annotation for Systems Interoperability,” (May 2014).

Lin, Y.C. *et al.* (2020) “Evaluating cross-lingual semantic annotation for medical forms,” *HEALTHINF 2020 - 13th International Conference on Health Informatics, Proceedings; Part of 13th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2020*, (Biostec), pp. 145–155. doi:10.5220/0008979901450155.

Lin, Y.C., Hoffmann, P. and Rahm, E. (2021) “Enhancing cross-lingual semantic annotations using deep network sentence embeddings,” in *HEALTHINF 2021 - 14th International Conference on Health Informatics; Part of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2021*. SciTePress, pp. 188–199. doi:10.5220/0010256801880199.

Marovac, U. and Avdic, A. (2021) “The Tools and Resources for Clinical Text Processing,” in. Singidunum University, pp. 134–140. doi:10.15308/sinteza-2021-134-140.

Martínez-romero, M. *et al.* (2017) “NCBO Ontology Recommender 2 . 0 : an enhanced approach for biomedical ontology recommendation,” pp. 1–22. doi:10.1186/s13326-017-0128-y.

Oliveira, P. and Rocha, J. (2013) “Semantic annotation tools survey,” *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, (March 2015), pp. 301–307. doi:10.1109/CIDM.2013.6597251.

Omid Yousefianzadeh, A.T. (2020) “COVID-19 Ontologies and their Applications in Medical Sciences: Reviewing BioPortal,” pp. 30–35.

Özgür, A., Hur, J. and He, Y. (2016) “The Interaction Network Ontology-supported modeling and mining of complex interactions represented with multiple keywords in biomedical literature,” *BioData Mining*, 9(1), pp. 1–18. doi:10.1186/s13040-016-0118-0.

Perera, N., Dehmer, M. and Emmert-Streib, F. (2020) “Named Entity Recognition and Relation Detection for Biomedical Information Extraction,” *Frontiers in Cell and Developmental Biology*. Frontiers Media S.A. doi:10.3389/fcell.2020.00673.

Perez, N. *et al.* (2020) “Cross-lingual semantic annotation of biomedical literature: Experiments in Spanish and English,” *Bioinformatics*, 36(6), pp. 1872–1880. doi:10.1093/bioinformatics/btz853.

Savova, G.K. *et al.* (2010) “Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications,” *Journal of the American Medical Informatics Association*, 17(5), pp. 507–513. doi:10.1136/jamia.2009.001560.

Shah, N.H. *et al.* (2009) “Comparison of concept recognizers for building the open biomedical annotator,” *BMC Bioinformatics*, 10(SUPPL. 9), pp. 1–9. doi:10.1186/1471-2105-10-S9-S14.

Sioutos, N. *et al.* (2007) “NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information,” *Journal of Biomedical Informatics*, 40(1), pp. 30–43. doi:10.1016/j.jbi.2006.02.013.

Sreeninvasan, M. and Chacko, A. (2020) “A Case for Semantic Annotation of EHR,” *Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020*, pp. 1363–1367. doi:10.1109/COMPSAC48688.2020.00-66.

Stewart, S.A., von Maltzahn, M.E. and Abidi, S.S.R. (2012) “Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons,” *CEUR Workshop Proceedings*, 895, pp. 63–77.

le Sueur, H. *et al.* (2020) “The challenges in data integration - Heterogeneity and complexity in clinical trials and patient registries of Systemic Lupus Erythematosus,” *BMC Medical Research Methodology*, 20(1), pp. 1–5. doi:10.1186/s12874-020-01057-0.

Tanenblatt, M., Coden, A. and Sominsky, I. (2010) “The ConceptMapper approach to named entity recognition,” *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, (January), pp. 546–551.

Tchechmedjiev, A. *et al.* (2018) “SIFR annotator: Ontology-based semantic annotation of French biomedical text and clinical notes,” *BMC Bioinformatics*, 19(1). doi:10.1186/s12859-018-2429-2.

Teng, R. and Verspoor, K. (2017) *READ-Biomed-Server: A Scalable Annotation Server Using the UIMA Concept Mapper*. Available at: http://geneontology.org/page/download-ontology.

Tseytlin, E. *et al.* (2016) “NOBLE - Flexible concept recognition for large-scale biomedical natural language processing,” *BMC Bioinformatics*, 17(1), pp. 1–15. doi:10.1186/s12859-015-0871-y.

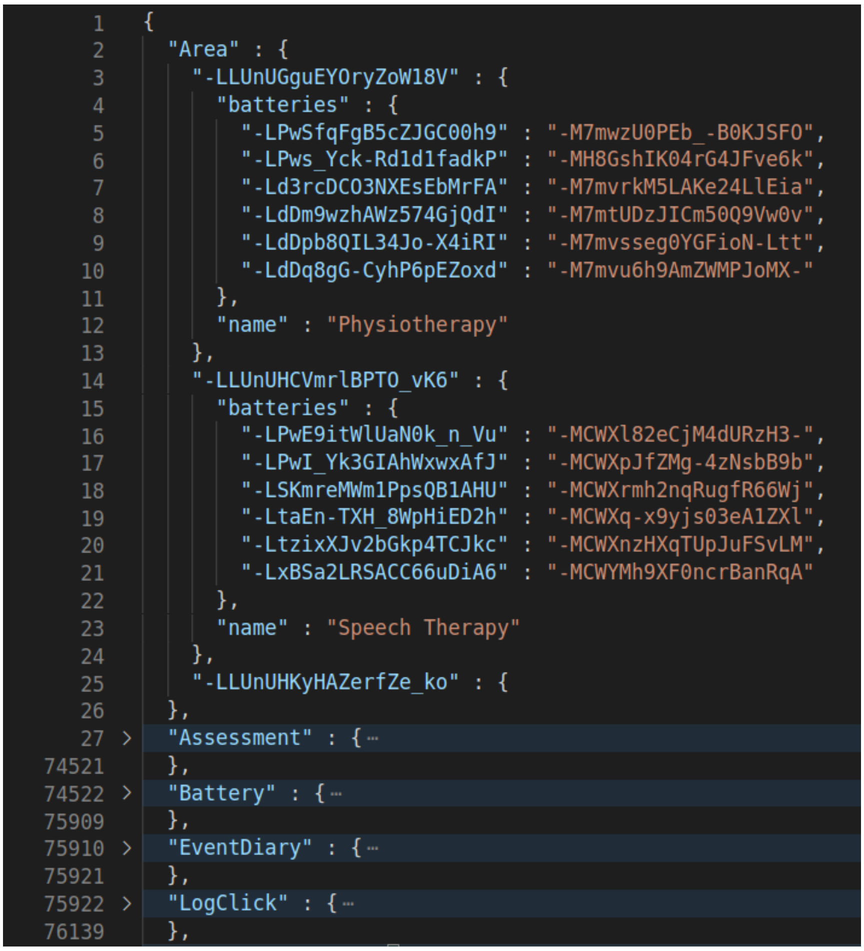
Vidal, M.E. *et al.* (2019) *Semantic data integration of big biomedical data for supporting personalised medicine*, *Studies in Computational Intelligence*. doi:10.1007/978-3-030-06149-4_2.

Yadav, V. and Bethard, S. (2019) “A Survey on Recent Advances in Named Entity Recognition from Deep Learning models.” Available at: http://arxiv.org/abs/1910.11470.

Zapata-Ospina, J.P. and García-Valencia, J. (2020) “Validity based on content: A challenge in health measurement scales,” *Journal of Health Psychology* [Preprint]. SAGE Publications Ltd. doi:10.1177/1359105320953477.

Appendix A

Extracted Json DataPark DataBase



```
{
  "Area" : {
    "-LLUnUGguEYOryZoW18V" : {
      "batteries" : {
        "-LPwSfqFgB5cZJGC00h9" : "-M7mwzU0PEb_-B0KJSFO",
        "-LPws_Yck-Rd1d1fadkP" : "-MH8GshIK04rG4JFve6k",
        "-Ld3rcDCO3NXEsEbMrFA" : "-M7mvrkM5LAKe24LlEia",
        "-LdDm9wzhAWz574GjQdI" : "-M7mtUDzJICm50Q9Vw0v",
        "-LdDpb8QIL34Jo-X4iRI" : "-M7mvsseg0YGFioN-Ltt",
        "-LdDq8gG-CyhP6pEZoxd" : "-M7mvu6h9AmZWMPJoMX-"
      },
      "name" : "Physiotherapy"
    },
    "-LLUnUHCVmrlBPTO_vK6" : {
      "batteries" : {
        "-LPwE9itWlUaN0k_n_Vu" : "-MCWXl82eCjM4dURzH3-",
        "-LPwI_Yk3GIAhWxwxAfJ" : "-MCWXpJfZMg-4zNsbB9b",
        "-LSKmreMWm1PpsQB1AHU" : "-MCWXrmh2nqRugfR66Wj",
        "-LtaEn-TXH_8WpHiED2h" : "-MCWXq-x9yjs03eA1ZXl",
        "-LtzixXJv2bGkp4TCJkc" : "-MCWXnzHXqTUpJuFSvLM",
        "-LxBSa2LRSACC66uDiA6" : "-MCWYMh9XF0ncrBanRqA"
      },
      "name" : "Speech Therapy"
    },
    "-LLUnUHKyHAZerfZe_ko" : {
  },
  "Assessment" : {…
  },
  "Battery" : {…
  },
  "EventDiary" : {…
  },
  "LogClick" : {…
  },
```

Figure 28: Organization of Json file hierarchy.