Faculdade de Engenharia da Universidade do Porto

Indexação de Documentos Clínicos

João de Sá Balão Calisto Correia

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Gabriel David (Prof. Associado, FEUP)

Co-Orientador: Francisco Correia (Engº, Glintt - Healthcare Solutions)

Fevereiro de 2016

© João de Sá Balão Calisto Correia, 2016

Resumo

Hoje em dia o acesso a informação estruturada e bem organizada é cada vez mais essencial no planeamento, desenvolvimento e desempenho de uma empresa. No meio clínico o acesso a este tipo de informação influencia os serviços prestados pelas entidades médicas, uma vez que a informação encontra-se demasiado dispersa por diversas fontes. Torna-se assim claro que inovar o acesso à informação nestes serviços é essencial para melhorar a eficiência e qualidade dos mesmos.

Na perspetiva de melhorar estes serviços, surge a solução de indexação de documentos clínicos, que possibilita a extração de diversos documentos com diversos formatos de uma base de dados da Glintt, e indexa e disponibiliza a sua informação numa área de pesquisa. Os formatos em causa podem ser PDF, Word, XML, JSON entre outros. Relativamente ao conteúdo da informação temos por exemplo: resultados analíticos laboratoriais, relatórios clínicos, diagnósticos codificados em ICD-9, notas clínicas dos médicos, requisições de exames, prescrições de medicamentos e informação demográfica de pacientes. Como forma de reduzir os pedidos à base de dados em causa e, para uma maior uniformização da informação segue-se a norma EHR enquanto que, para fins de indexação, utiliza-se o Solr.

Este projeto destina-se a profissionais de saúde e, por isso, disponibiliza uma forma centralizada e pesquisável de obter todos os dados clínicos de um doente/paciente assim como a informação documental correspondente.

Dito isto, a solução pretende causar um forte impacto no dia-a-dia dos profissionais de saúde assim como nos seus pacientes, tornando o acesso à informação clínica mais rápido e simples.

Abstract

Nowadays access to structured and well organized information is increasingly important in the planning, development and performance of a company. In a clinical environment, the access to this type of information influences the services provided by the medical entities, since the information is too dispersed by various sources. Thus it becomes clear that innovating access to information on these services is essential to improve efficiency and quality.

In the perspective of improving these services, there is the indexation of clinical documents solution, which enables the extraction of several documents in various formats of a Glintt database, and index, and provides its information in a search area. The formats in question can be PDF, Word, XML, JSON and more. Regarding the content of the information we have for example: laboratory analytical results, clinical reports, diagnoses coded in ICD-9, clinical notes from doctors, requisitions, drug prescriptions and demographic information of patients. In order to reduce the requests to the database in question and, to a greater standardization of information follows the EHR standard while, for indexing purposes, we use Solr.

This project is intended to health professionals and therefore provides a centralized, searchable way to get all the clinical data of a patients and the corresponding documentary information.

That said, the solution aims to have a strong impact on the day-to-day health professionals as well as their patients, making access easier and faster clinical information.

Conteúdo

[Introdução 1](#_Toc452999206)

[1.1 Contexto/Enquadramento 1](#_Toc452999207)

[1.2 Motivação e Objetivos 2](#_Toc452999208)

[1.3 Estrutura do Relatório 2](#_Toc452999209)

[Revisão Bibliográfica 3](#_Toc452999210)

[2.1 Introdução 3](#_Toc452999211)

[2.2 Recuperação de Informação 4](#_Toc452999212)

[2.2.1 Modelos Clássicos 5](#_Toc452999213)

[2.2.2 Modelos Estruturados 6](#_Toc452999214)

[2.3 Indexação 7](#_Toc452999215)

[2.3.1 Apache Lucene 9](#_Toc452999216)

[2.3.2 Apache Solr 11](#_Toc452999217)

[2.3.3 ElasticSearch 11](#_Toc452999218)

[2.3.4 Sphinx Search 12](#_Toc452999219)

[2.3.5 Análise Tecnológica 13](#_Toc452999220)

[2.4 Classificação 15](#_Toc452999221)

[2.4.1 Relações Hierárquicas 16](#_Toc452999222)

[2.4.2 UMLS 16](#_Toc452999223)

[2.5 EHR 19](#_Toc452999224)

[2.5.1 AllScripts 21](#_Toc452999225)

[2.5.2 EHRserver (cabolabs) 21](#_Toc452999226)

[2.6 Resumo 22](#_Toc452999227)

[Solução 23](#_Toc452999228)

[3.1 Arquitetura 23](#_Toc452999229)

[3.1.1 E-results (PL/SQL) 24](#_Toc452999230)

[3.1.2 Módulo de Mapeamento 25](#_Toc452999231)

[3.1.3 EHR (MySQL) 25](#_Toc452999232)

[3.1.4 Módulo de Indexação 25](#_Toc452999233)

[3.1.5 Índices 25](#_Toc452999234)

[3.1.6 Módulo de Pesquisa 26](#_Toc452999235)

[3.1.7 Interface 26](#_Toc452999236)

[3.2 Diagrama de casos de uso 26](#_Toc452999237)

[3.3 Conclusões 27](#_Toc452999238)

[Implementação 29](#_Toc452999239)

[4.1 Protótipo 29](#_Toc452999240)

[4.2 Bases de dados da aplicação 29](#_Toc452999241)

[4.2.1 E-results 29](#_Toc452999242)

[4.2.2 EHR 30](#_Toc452999243)

[4.3 Mapeamento 30](#_Toc452999244)

[4.4 Indexação 32](#_Toc452999245)

[4.5 Pesquisa 35](#_Toc452999246)

[4.6 Interface 35](#_Toc452999247)

[4.7 Resumo 35](#_Toc452999248)

[4.7.1 Problemas 35](#_Toc452999249)

[4.7.2 Software 35](#_Toc452999250)

[4.7.3 Conclusão 35](#_Toc452999251)

[Simulação e Resultados 37](#_Toc452999252)

[5.1 Cenário 37](#_Toc452999253)

[5.2 Resultados Experimentais 37](#_Toc452999254)

[Conclusões e Trabalho Futuro 39](#_Toc452999255)

[6.1 Satisfação dos Objetivos 39](#_Toc452999256)

[6.2 Trabalho Futuro 40](#_Toc452999257)

[Referências 41](#_Toc452999258)

[Anexo A 42](#_Toc452999259)

Lista de Figuras

[Figura 1 - Sistema RI 4](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999188)

[Figura 2 - Arquitetura do Apache Lucene, Hatcher and Gospodnetic [4] 10](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999189)

[Figura 3 - Arquitetura do *Sphinx Search* [2] 12](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999190)

[Figura 4 - Ranking de popularidade de ferramentas de pesquisa e bases de dados [3] 15](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999191)

[Figura 5 - Estrutura hierárquica e relações dos termos do SNOMED [5] 17](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999192)

[Figura 6 - Estrutura hierárquica do MeSH [1] 19](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999193)

[Figura 7 - Atores envolvidos no *openEHR*, bibliografia 21](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999194)

[Figura 8 - Estrutura da aplicação cliente/servidor 23](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999195)

[Figura 9 - Arquitetura da solução 24](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999196)

[Figura 10 - Diagrama de casos de uso do utilizador 27](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999197)

[Figura 11 - *Template* (dados demográficos) 31](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999198)

[Figura 12 - Planeamento 40](file:///C:\Users\Joaogcorreia\Desktop\EHR%20+%20Solr%20+%20IndexDocClinicos\Indexacao-de-Documentos-Clinicos\Dissertacao\Dissertação_final.docx#_Toc452999199)

Lista de Tabelas

Tabela 1 - Comparação entre SOLR, ElasticSearch e Sphinx Search 14

Tabela 2 - Exemplo de ICD-9-CM [21] 18

Tabela 3 - Modelo de referência, bibliografia 20

Tabela 4 - Parâmetros de indexação do *Solr* 34

Abreviaturas e Símbolos

|  |  |
| --- | --- |
| API | Application Programming Interface |
| EER | Extended Entity Relationship |
| EHR | Electronic Health Record |
| HTTP | Hypertext Transfer Protocol |
| ICD-9-CM | International Classification of Diseases, 9th Revision, Clinical Modification |
| JSON | JavaScript Object Notation |
| MeSH | Medical Subject Headings |
| MySQL | My Structured Query Language |
| PDF | Portable Document Format |
| REST | Representational State Transfer |
| RI | Recuperação de Informação |
| SNOMED CT | Systematized Nomenclature of Medicine – Clinical Terms |
| UMLS | Unified Medical Language System |
| XML | eXtensible Markup Language |

Capítulo 1

# Introdução

Neste capítulo introdutório apresenta-se o contexto da dissertação, a motivação e os objetivos mostrando também as razões pelas quais o tema em causa é importante e que influencia tem/terá no dia-a-dia dos seus utilizadores. É também apresentada uma visão global sobre a estrutura da mesma.

## Contexto/Enquadramento

Com o passar dos anos, constata-se que a sociedade em que vivemos é baseada cada vez mais na informação. A quantidade de informação aumenta a um ritmo crescente, o que torna a sua gestão bastante difícil. Assim, para que a sociedade se desenvolva, torna-se progressivamente necessária a existência de soluções e mecanismos inovadores que possibilitem o acesso à informação de forma rápida e eficaz. Para fazer face a estes factos, surgem assim métodos de recuperação e de indexação que tornam possível o acesso à informação com as características referidas anteriormente.

Assim sendo, é possível assumir que esta dissertação se encontra nas áreas das ciências da informação e nas tecnologias de bases de dados, na medida em que aborda a importância do acesso e recuperação da informação em meios clínicos.

Esta dissertação será desenvolvida no âmbito empresarial da *Glintt - Healthcare Solutions*, a qual é focada no ramo da saúde e onde, a nível nacional, é líder destacada neste segmento do mercado. A sua sede está situada na cidade do Porto e é constituída por cerca de 300 colaboradores. A *Glintt* trabalha com uma enorme quantidade de informação clínica que pode variar desde documentos PDF, Word, JSON a informação distribuída pelas tabelas das bases de dados, sendo por isso necessário uma boa organização e tratamento da mesma.

## Motivação e Objetivos

No meio clínico a quantidade de informação é elevada, o que torna difícil o acesso à mesma pelas entidades médicas. Além disto, outros fatores alarmantes são que a mesma quantidade de informação encontra-se dispersa por várias fontes e diferentes formatos, e por vezes, não se encontra estruturada o que faz com que as entidades médicas percam demasiado tempo na pesquisa/ procura dos dados clínicos de um doente. Relativamente aos diferentes formatos e fontes em que a informação se encontra, tanto se pode falar de documentos como de campos de bases de dados em texto livre, sendo exemplos destes: resultados analíticos laboratoriais, relatórios clínicos, diagnósticos codificados em ICD-9, notas clínicas dos médicos, requisições de exames, prescrições de medicamentos e informação demográfica de pacientes.

Devido a esta diversidade na origem e tipos dos dados, há um atraso no acesso à informação por parte dos médicos, há falta de informação e, por vezes, esta nem chega corretamente às mãos das entidades clínicas. Assim sendo, está clara a necessidade de organizar e tratar esta informação. Porém fazê-lo não é tão trivial como possa parecer pois, apesar da crescente atualização dos documentos clínicos para formatos eletrónicos, estes muitas vezes resultam apenas numa digitalização do documento original e, por isso, surgem diversos problemas que levantam algumas questões: Como indexar imagens? Como pesquisar algo sobre elas? Outro aspeto é também a forma de apresentar os resultados da pesquisa, atendendo a que as fontes podem estar em documentos autónomos ou como valores em campos de bases de dados. Estes últimos pontos realçam a grande variedade de dados e, com isso, alguns dos problemas que são necessários resolver nesta área.

Com o fim de resolver estes problemas surge a necessidade de classificar e indexar a informação que tornam possível uma uniformização dos dados numa só estrutura, permitindo assim, que estes sejam disponibilizados numa plataforma única e pesquisável. Em suma, o acesso à informação torna-se mais fácil e de qualidade, uma vez que toda a informação se encontra num só ponto.

## Estrutura do Relatório

O presente documento está dividido em 3 capítulos:

* **Capítulo 1 – Introdução**:é um capítulo onde é apresentado o contexto ou enquadramento do tema em causa;
* **Capítulo 2 – Revisão Bibliográfica**: é onde se descreve o estudo do estado da arte e está dividido em três subsecções: Recuperação de Informação, Indexação e Classificação;
* **Capítulo 3 – Conclusões e Trabalhos futuros**: é onde se apresenta conclusões relativas a este documento e se revela as etapas futuras do projeto em causa.

Capítulo 2

# Revisão Bibliográfica

As secções seguintes descrevem o estado da arte relativa aos processos de RI (recuperação de informação) e indexação com uma breve descrição das tecnologias mais recentes e mais utilizadas na área. Na última secção é apresentada como é feita a classificação da informação.

## Introdução

Desde há muito tempo atrás que a comunidade tem vindo a armazenar grandes quantidades de informação para que esta possa ser consultada por diversas pessoas de geração em geração. Face a esta acumulação de informação recolher/extrair informação com significado/utilidade começou a tornar-se bastante complicado. Entre 1948 e 1950, Calvin Mooers [6] cria o termo “Recuperação de Informação” pressionado pela necessidade de gerir a explosão da informação na literatura científica na segunda metade do século XX.

Nos anos seguintes foram feitas várias experiências na área e durante os anos 70 os sistemas de recuperação de informação começaram a tornar-se reais essencialmente devido ao aparecimento de processadores de texto [7].

Em 1980 com a diminuição constante do preço do espaço em discos a informação disponível em máquinas crescia rapidamente. Nesta década começaram a estar disponíveis textos completos e não apenas resumos e índices.

Com o passar dos anos e, na sequência do aparecimento de diversos dispositivos para armazenamento de informação digital, a informação continuou a crescer forçando os sistemas de recuperação de informação a desenvolverem-se da mesma forma com o objetivo de acompanhar o aumento de informação.

Hoje em dia, o processo de recuperação de informação está bastante disseminado na sociedade, sendo que o desenvolvimento de ferramentas de pesquisa veio ajudar neste processo. O conceito de recuperação de informaçãosurgiu então da forte necessidade de extrair informação com valor das diversas fontes, organizando e facilitando o acesso aos conteúdos informacionais através dos processos de indexação e de descrição.

## Recuperação de Informação

A RI pode ser definida como a atividade de obtenção de informação, devido a uma necessidade, a partir de uma coleção de recursos de informação. Segundo alguns autores a RI é descrita da seguinte forma:

“*Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).*”, Manning, et al. [8]

De uma forma geral, os sistemas RI devem revelar ao utilizador resultados interessantes e com valor para o mesmo, recorrendo, para isso, a um processo de ordenação da informação encontrada após a pesquisa. Este processo é realizado tendo em conta a pesquisa feita pelo utilizador, ou seja, resultados mais próximos dos termos pesquisados deverão aparecer nas primeiras posições, enquanto que resultados mais distantes deverão aparecer nas últimas posições da lista de resultados ordenada.



Figura - Sistema RI

Na imagem anterior é ilustrado o funcionamento de um sistema RI. Todavia, a estratégia apresentada na figura 1 não é tão trivial como possa parecer, uma vez que um sistema RI enfrenta vários problemas, entre os quais se destaca o facto de não se saber quais os documentos relevantes para determinada pergunta do utilizador. Esta tarefa é normalmente realizada com base em alguma heurística que decide quais os documentos com maior relevância a serem recuperados e ordena-os a partir de critérios estabelecidos [9].

Para a resolução deste problema surgem modelos para recuperação de informação. Estes estão divididos em modelos clássicos e modelos estruturados. Os modelos clássicos [6] são baseados em termos de indexação (palavras-chave), ou seja, cada documento consiste num conjunto desses termos. Já os modelos estruturados podem especificar alguma informação sobre a estrutura do texto, por exemplo, proximidade de palavras. A descrição de cada um dos modelos referidos encontra-se nos tópicos seguintes.

Em relação aos sistemas RI apenas resta dizer que têm um papel muito importante ou até fundamental na satisfação das necessidades do utilizador e por isso, adaptar um sistema deste tipo num meio clínico tem como objetivo ajudar, neste caso, as entidades médicas, de modo a que estas melhorem o seu desempenho no dia-a-dia.

### Modelos Clássicos

**Modelo Booleano**

É um modelo clássico baseado na lógica booleana, isto é, os termos pretendidos são combinados com os operadores AND, OR, e NOTde forma a fornecer ao utilizador informação mais específica, por exemplo: “presidente AND lincoln”. Oferece algumas vantagens como o facto de os resultados serem previsíveis e fáceis de explicar e, como é um conceito intuitivo, os utilizadores sentem-se no controlo do sistema. No entanto são também visíveis algumas desvantagens dentro das quais se destaca o facto de não providenciar uma ordenação dos documentos encontrados em resposta à pergunta do utilizador [10].

**Modelo Vetorial**

No modelo vetorial os documentos e as *queries* são representados por vetores de termos que têm um peso associado, por exemplo: [(palavra1, peso1), (palavra2, peso2), … (palavra n, peso n)]. Este peso não é binário de forma a não limitar os resultados da pesquisa (como é feito no modelo booleano) e é obtido através da quantidade de vezes que um termo aparece no próprio documento e nos outros documentos. Os pesos da *querie* e do documento são calculados através do peso dos respetivos vetores. Uma vez efetuado este cálculo parte-se para o grau de similaridade que pode ser obtido com base no ângulo entre os dois vetores referidos (*query* e documento) [11].

A similaridade permite a saída de resultados ordenados, mas com esta abordagem um documento relevante pode não conter os termos pesquisados pelo utilizador. Isto acontece essencialmente devido ao facto de o modelo assumir que os termos são independentes. Assim sendo, pode-se dizer que o modelo encontra apenas resultados parciais face a uma pesquisa.

**Modelo Probabilístico**

O modelo probabilístico foi proposto por S.E. Robertson e K. Sparck Jones [12], e, de uma forma geral, consiste em estimar a probabilidade de um documento ser relevante para uma *query* q. Este modelo assume que a probabilidade de relevância depende apenas da *query* e do documento. O resultado ideal de uma pesquisa seria um conjunto que contivesse todos e apenas os documentos relevantes para o utilizador. Este modelo ordena os documentos na ordem decrescente da probabilidade de relevância da informação.

Existem algumas limitações, entre as quais se verifica a ausência de atribuição de pesos a cada termo. Além disto, é necessário fazer uma estimativa inicial de forma a refinar o conjunto da resposta ideal sendo que, para cada termo, não é considerada a quantidade de vezes que este aparece num documento.

### Modelos Estruturados

**Modelo da Proximidade de Nós**

O modelo da proximidade de nós visa a recuperação de documentos através de uma estrutura hierárquica de índices e, por isso, permite que várias estruturas possam ser definidas no mesmo texto, cada uma sendo uma hierarquia restrita mas permitindo a sobreposição entre áreas delimitadas por hierarquias diferentes. Uma *query* pode relacionar diferentes hierarquias mas retorna um subconjunto de nós de apenas uma delas. Cada nó tem associado um segmento que corresponde à área de texto a que este está associado. O segmento de um nó inclui os dos seus descendentes.

De uma forma geral, o modelo em causa possibilita a definição de estruturas de indexação hierárquicas e independentes sobre um mesmo documento [13].

**Modelo de Listas Não Sobrepostas**

O modelo de listas não sobrepostas foi proposto por Burkowski [9] em 1992 e consiste na divisão de todo o texto de um documento em regiões não sobrepostas e juntá-las numa lista. Existem diversas formas de dividir o texto em regiões não sobrepostas que levam à geração de listas múltiplas:

* Lista para capítulos;
* Lista para secções;
* Lista para subsecções.

Regiões de texto de listas distintas podem sobrepor-se.

Relativamente à implementação, o modelo utiliza um ficheiro invertido que combina as palavras-chave e as regiões de texto. Para cada entrada neste ficheiro invertido é associada uma lista de regiões de texto. Listas de regiões de texto podem ser juntadas com listas de palavras-chave.

## Indexação

Como referido no capítulo anterior, os sistemas RI organizam e viabilizam o acesso à informação através da descrição de documentos e da operação do tratamento documental, a indexação. Mas afinal, em que consiste a indexação?

A indexação tenta descrever e caracterizar um documento com o auxílio de representações dos conceitos contidos nesses documentos, ou seja, em transcrever para linguagens de indexação os conceitos depois de terem sido extraídos dos documentos por meio de uma análise dos mesmos. Este processo de indexar segue assim três etapas:

1. Compreensão do documento;
2. Seleção dos conceitos para a pesquisa;
3. Tradução destes conceitos nas linguagens de indexação.

Na compreensão do documento deve ser feita uma análise global do mesmo, isto é, ter em consideração as partes mais úteis como fonte de informação do tema/assunto. As partes referidas podem ser: título, resumo, índice, prefácio, introdução, entre outros. Nesta etapa não é, por isso, necessário ler o documento na íntegra.

Relativamente à segunda etapa, a seleção dos conceitos para a pesquisa, deve ser feita de forma exaustiva e específica, ou seja, todos os assuntos de que trata o documento devem ser registados e nunca um conceito deve ser traduzido.

No terceiro passo, é feita a tradução destes conceitos nas linguagens de indexação, isto é, transformar os conceitos selecionados em termos ou símbolos autorizados para representá-los no sistema. Após a execução destas etapas dá-se por terminado o processo de indexação.

Neste processo de indexar é ainda importante realçar dois conceitos: *stemming* e *stopwords*.

***Stemming***

No processo de indexar e, na maioria das vezes, existem palavras que, apesar de pertencerem à mesma família, são indexadas separadamente. Como é possível observar, nestes casos estão a ser criados vários índices que possivelmente representarão o mesmo conteúdo. Face a este e outros casos semelhantes surge o *stemming* [14] que consiste em reduzir as palavras à sua raiz morfológica. A raiz morfológica de uma palavra é o conjunto de caracteres que está presente em todas as suas derivações.

Pode parecer um processo simples, mas, se o fosse, apenas existiria uma única implementação dele, o que não é o caso. O *stemming* por vezes falha em reduzir as palavras com o mesmo significado para a mesma raiz (*understemming*) e por outras vezes falha em manter palavras com significados diferentes separadas (*overstemming*). Apesar disto, um algoritmo de *stemming* bem implementado é capaz de melhorar significativamente a qualidade da solução.

***Stopwords***

Em todos os textos/documentos existem palavras que, por aparecerem em muitos documentos, não acrescentam valor discriminatório relativo ao conteúdo de um documento como o “e”, “de”, etc. Além disso, estas palavras são raramente pesquisadas pelo utilizador uma vez que não têm valor para o documento. As palavras referidas são conhecidas como *stopwords*, pois devido à sua falta de valor contextual são eliminadas do processo de indexação. Este processo de eliminação referido permite um aumento de qualidade nos resultados.

Nos tópicos seguintes são apresentadas algumas das ferramentas mais usadas nos dias de hoje para indexar e pesquisar conteúdos de uma forma mais rápida e fácil. Essas ferramentas são: *Apache Lucene*, *Apache Solr*, *ElasticSearch* e *SphinxSearch*.

### Apache Lucene

O *Apache Lucene* foi desenvolvido em 1997-8 por *Doug Cutting*. Inicialmente denominado apenas de *Lucene*, foi disponibilizado gratuitamente no *SourceForge* em 2000. Mais tarde, em 2001, juntou-se à *Apache Software Foundation's* onde passou a chamar-se *Apache Lucene*. Em 2005, já era um projeto de alto nível na *Apache Software Foundation's*. Segundo os autores Hatcher e McCandless o *Lucene* é descrito da seguinte forma:

“*Lucene is a high-performance, scalable information retrieval (IR) library. IR refers to the process of searching for documents, information within documents, or metadata about documents. Lucene lets you add searching capabilities to your applications. It’s a mature, free, open source project implemented in Java, and a project in the Apache Software Foundation, licensed under the liberal Apache Software License. As such, Lucene is currently, and has been for quite a few years, the most popular free IR library.*”, McCandless, et al. [15]

Descrito em outras palavras, *Lucene* é uma ferramenta que permite indexar e realizar pesquisa sobre os elementos indexados. Um índice é constituído por uma lista de documentos (ou outra fonte de informação), os quais são compostos por uma sequência de campos e em que cada um destes consiste numa sequência nomeada de termos, ou seja, por nome do campo e termo, por exemplo: (“título”, ”introduction to Lucene”). O termo pode conter uma ou várias palavras com valor no contexto do documento a indexar. O *Lucene* segue um sistema de indexação designado por *inverted índex*, ou em português, índice invertido, pois em vez de, para cada documento, guardar os termos que nele existem, guarda, para cada termo, os documentos onde estes se encontram.

Para determinar a relevância de um documento perante a pergunta do utilizador, o *Apache* *Lucene* utiliza uma combinação entre o modelo booleano e o modelo vetorial.

A **figura 2** apresenta a arquitetura do *Apache Lucene* em 2004, apesar de continuar bastante atual.



Figura - Arquitetura do Apache Lucene, Hatcher and Gospodnetic [4]

* **Gather Data:** nesta etapa são recolhidos os conteúdos dos diferentes documentos.
* **Index Documents:** é feita uma análise do documento e, após esta, inicia-se o processo de indexação.
* **Index:** base de dados de índices.
* **Get Users’ Query:** uma vez feito o pedido de pesquisa do utilizador, a aplicação constrói a *query* com base no texto de pesquisa a fim de ser utilizada na interrogação às bases de dados de índices.
* **Search Index:** é a etapa onde se realiza a pesquisa por índices de acordo com a *query* construída a partir dos dados de pesquisa introduzidos pelo utilizador.
* **Present Search Results:** apresenta o resultado da pesquisa ao utilizador.

Com isto, termina a análise desta ferramenta afirmando que é uma ferramenta poderosa de indexação e pesquisa que permite rapidamente e eficazmente criar índices e realizar pesquisa sobre os mesmos.

### McCandless, Hatcher [1]Apache Solr

O *Apache Solr* surgiu em 2006 e consiste numa ferramenta construída a partir do *Apache Lucene*. Devido à sua origem a partir do *Apache Lucene*, implementa muitas das suas funcionalidades e acrescenta muitas outras. O *Apache Solr* apresenta assim as seguintes características [16]:

* Altamente escalável e tolerante a falhas;
* Otimizado para alto volume de tráfego;
* Fornece APIs baseadas em REST o que lhe permite ser integrado a praticamente qualquer linguagem de programação (XML, HTTP e JSON APIs);
* Fácil extensão através de *plugins*;
* Quase em tempo real, isto é, os documentos estão disponíveis para pesquisa quase imediatamente a seguir à sua indexação;
* Integra o *Apache Tika* [17], por isso, aceita para indexação documentos com diversas extensões como PDF, Word, XML, etc.

Além destas características pode ser ainda comentado o facto de este possuir uma boa documentação [18] que, juntamente com o excelente suporte que recebe, o torna uma ferramenta bastante utilizada hoje em dia.

### ElasticSearch

Em semelhança ao *Apache Solr*, o *ElasticSearch* foi desenvolvido com base no *Apache Lucene* ao qual foram adicionadas algumas melhorias. Foi desenvolvido por Shay Banon e teve o primeiro lançamento em 2010.

O *ElasticSearch* permite facilmente aceder às funcionalidades do *Lucene* para indexação e pesquisa [19]. No entanto, proporciona um novo nível de abstração de análise em tempo real. Outro nível de abstração é a forma como podem ser organizados os documentos: múltiplos índices podem ser pesquisados em simultâneo ou separadamente, e é possível colocar diferentes tipos de documentos dentro de cada índice.

Por fim o *ElasticSearch*, como o próprio nome sugere, é elástico, ou seja, é possível acrescentar gradualmente servidores de forma a aumentar a capacidade e tolerância a falhas. De forma semelhante, é possível remover estes gradualmente de forma a reduzir os custos.

### Sphinx Search

O *Sphinx Search* começou a ser desenvolvido em 2001 e consiste num servidor de pesquisa *full-text*, escrito em C++ que funciona com diversos sistemas operativos. Esta ferramenta permite indexar e pesquisar informação guardada em bases de dados SQL e NoSQL, ou apenas em ficheiros.

Na figura seguinte é apresentada a arquitetura da ferramenta.



Figura - Arquitetura do *Sphinx Search* [2]

Descrevendo com melhor pormenor o que se observa na imagem, o **searchd** é responsável por receber pedidos do cliente e executar pesquisa comparando com os índices no **Sphinx**. Já o **Indexador Sphinx** é responsável por recolher a informação do **MySQL** (ou outra fonte suportada) e indexá-la.

Perante esta análise é ainda possível realçar algumas características [2] desta ferramenta:

* Indexação em tempo real;
* Indexação de base de dados SQL (MySQL, PostgreSQL, Oracle, SQLite, etc);
* Indexação de base de dados NoSQL;
* Vem com o *SphinxAPI* que é uma biblioteca nativa disponível para Java, PHP, Python, Perl, C entre outras linguagens;
* Suporta pesquisas sintaticamente complexas;
* Analisa proximidade entre palavras e ordena em rankings.

### Análise Tecnológica

Nos tópicos anteriores foram descritas algumas das ferramentas usadas hoje em dia para efeitos de indexação e recuperação de informação. Não esquecendo que estamos perante uma indexação de documentos clínicos, é importante ter em conta quais das tecnologias melhor respondem às necessidades em causa. Para tal, e relembrando o que está em causa, é necessário recolher e extrair informação de diferentes fontes e com diferentes formatos. As fontes possíveis são: *webservices*, sistemas de ficheiros, bases de dados relacionais e *document-oriented*. Relativamente aos formatos podem variar desde ficheiros de texto simples a PDF, Word, JSON, XML entre outros. Face a estes dados é possível partir para uma análise comparativa entre as ferramentas em estudo.

Relativamente ao *Apache Lucene*, é possível afirmar que não trará grandes vantagens na sua utilização direta no âmbito em questão, pois os seus “sucessores”, *Solr* e *ElasticSearch* trazem uma abrangência muito maior relativamente, não só, às linguagens que disponibilizam, mas também aos diferentes formatos nos quais pode ser realizada a indexação e extraída a informação. Perante esta situação, na tabela seguinte é feita uma análise comparativa das ferramentas *Apache Solr*, *ElasticSearch* e *Sphinx Search*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Solr** | | **ElasticSearch** | | **Sphinx Search** |
| API | | | | | |
| Bibliotecas | Ruby, Rails, PHP, Java, Python, Perl, C#, .Net, JavaScript | | Java, Groovy, PHP, Ruby, Perl, Python, .NET, JavaScript | | C++, Java, Perl, PHP, Python, Ruby |
| Indexação | | | | | |
| Importação de dados | JDBC, CSV, XML, Tika [17], URL, Flat File | | Amazon SQS, CouchDB, Dropbox, DynamoDB, FileSystem, Git, GitHub, Hazelcast, JDBC, JMS, Kafka, LDAP, MongoDB, neo4j, OAI, RabbitMQ, Redis, RSS, Sofa, Solr, St9, Subversion, Twitter, Wikipedia | | MySQL, PostgreSQL, MSSQL, ODBC source, XML pipe |
| Tempo Real |  | |  | |  |
| Licença | | | | | |
|  | | *Apache License* 2.0 | | *Apache License* 2.0 | GPLv2 e comercial |

Tabela - Comparação entre SOLR, ElasticSearch e Sphinx Search

Como é possível observar na tabela anterior, e começando por analisar a licença, tanto o *Solr* como o *Elasticsearch* têm a licença da Apache. Já o *Sphinx Search* tem a licença GPLv2 que, para fins comerciais, é necessário comprá-la. Face a esta informação temos aqui dois pontos de vista, ferramentas *opensource* e com licença comercial. Todavia ter licença comercial não é sinónimo de maior utilidade, isto é, no caso da importação de dados é possível observar que o *Sphinx Search* é muito mais orientado para as bases de dados e, por isso, carece de leitura de ficheiros PDF, Word entre outros que foram referidos anteriormente como uma necessidade para a solução em causa. Por outro lado, o *Solr* e o *Elasticsearch* implementam o *Apache Tika* [17], que consiste numa biblioteca que suporta a extração de texto de ficheiros binários, como por exemplo: PDF, Word, XML, formatos de áudio, etc. Mesmo relativamente às bibliotecas, o *Sphinx Search* apresenta menos variedade que as outras duas ferramentas. A figura seguinte apresenta um *ranking* de popularidade de diversas ferramentas de pesquisa e de bases de dados entre as quais se encontram as ferramentas estudadas neste projeto.



Figura - Ranking de popularidade de ferramentas de pesquisa e bases de dados [3]



Observando a figura anterior chegamos à conclusão que o *Solr* e *Elasticsearch* se encontram nas primeiras posições no que diz respeito a ferramentas de pesquisa e indexação. O *Sphinx Search* encontra-se na 35ª posição com uma popularidade de 9.08, bem inferior à do *Solr* e *Elasticsearch*. Perante esta informação e pelas características analisadas anteriormente segue-se a análise apenas com o *Solr* e *Elasticsearch*.

Decidir qual das duas ferramentas se adapta melhor ao problema em causa torna-se complicado, uma vez que as suas características são muito semelhantes. Segundo alguns autores tanto uma como outra merecem atenção:

“*Both search engines provide similar functionality, and features evolve quickly with each new version … the functionality you need is covered by both, and, as is often the case with competitors, choosing between them becomes a matter of taste.*”,Gheorghe, et al. [19]

Tendo em consideração a opinião dos autores, apenas resta apelar à documentação. No *ElasticSearch*, vê-se que há uma tentativa de produzir boa documentação, mas como ainda é uma ferramenta relativamente recente esta carece de algum conteúdo. Já a do *Solr* é bastante boa e a comunidade é bastante ativa o que leva a tender para esta ferramenta.

## Classificação

Com base no que foi dito até agora, é possível assumir que os resultados de uma pesquisa de elementos indexados seguem uma ordenação, ou não, por relevância, de acordo com os modelos clássicos e estruturados referidos anteriormente no tópico de recuperação de informação. Perante esta situação parece que o estudo levará a uma indexação simples de documentos sem uma perspetiva clínica associada, o que não é o caso. Pretende-se que a solução tenha teor clínico, isto é, que esteja ligada a uma classificação de termos, com um ponto de vista médico, de forma a apresentar ao utilizador resultados com mais significado e até mesmo acelerar o processo de pesquisa. A classificação referida consiste na ordenação sistemática de todos os conceitos numa área científica, neste caso, na área da medicina.

### Relações Hierárquicas

Segundo Rothwell, et al. [20], a relação mais importante usada para a classificação é uma relação hierárquica. Este tipo de classificação está dividida em:

* Classificação mono-hierárquica: um conceito só pode ter um pai na hierarquia;
* Classificação poly-hierárquica: um conceito pode ter vários pais na hierarquia.

Sendo assim, os autores referem ainda que uma mono-hierarquia não é suficiente para aplicações complexas, no entanto, as poly-hierarquias tendem a ser demasiado volumosas e redundantes não sendo, por isso, muitas vezes utilizadas. Para resolver tal problema os autores chegam à conclusão que a poly-hierárquia pode ser substituída por:

“Several disjoint monohierarchic classifications connected by a semantic model”, Rothwell, et al. [20]

Exemplo desta solução é o SNOMED, o qual se baseia num sistema de nomenclatura de termos médicos que providencia códigos, termos, sinónimos e definições usadas na documentação clínica.

### UMLS

O SNOMED é uma nomenclatura que faz parte das ferramentas do UMLS que consiste num conjunto de ficheiros e *software* que contemplam diversos vocabulários de saúde e biomedicina. O UMLS é composto por três ferramentas (fontes de conhecimento):

* ***Metathesaurus*** que consiste em termos e códigos de diversos vocabulários, incluindo CPT, ICD-10-CM, LOINC, MeSH, RxNorm, e SNOMED CT;
* ***Semantic Network*** que consiste em relações semânticas entre uma vasta gama de categorias;
* ***SPECIALIST Lexicon and Lexical Tools*** que consiste no processamento de linguagem natural.

As duas últimas ferramentas são utilizadas para construir o *Metathesaurus*. A construção deste envolve o processamento de termos e códigos usando as ferramentas lexicais, agrupando termos sinónimos em conceitos, categorizando conceitos semanticamente e incorporando relações e atributos providenciados pelos vocabulários. Nos tópicos seguintes são apresentadas as nomenclaturas SNOMED CT, MeSH e ICD-9.

**SNOMED CT**

Como já foi referido o SNOMED consiste num conjunto de termos organizados hierarquicamente e que têm relações entre si. A estrutura do SNOMED encontra-se na figura seguinte.



Figura - Estrutura hierárquica e relações dos termos do SNOMED [5]

**ICD-9**

O ICD-9 consiste num conjunto de códigos de diagnósticos e de procedimentos utilizados para classificação e codificação da informação de doenças e intervenções cirúrgicas, como é possível observar na tabela seguinte.

|  |  |
| --- | --- |
| Código | Descrição |
| 0010  0011  0019  0020 | Cholera due to vibrio cholerae  Cholera due to vibrio cholerae el tor  Cholera, unspecified  Typhoid fever |

Tabela - Exemplo de ICD-9-CM [21]

Fornece códigos relativos à classificação de doenças e de uma grande variedade de sinais, sintomas, aspetos anormais, queixas, circunstâncias sociais e causas externas para ferimentos ou doenças. A cada estado de saúde é atribuída uma categoria única à qual corresponde um código, que contém até 6 caracteres. Tais categorias podem incluir um conjunto de doenças semelhantes. Esta nomenclatura começou a ser desenvolvida do século 19 e é mantida pela organização mundial de saúde. A revisão mais conhecida é o ICD-9 que surgiu em 1975. (<http://portalcodgdh.min-saude.pt/index.php/Classifica%C3%A7%C3%A3o_Internacional_de_Doen%C3%A7as_(CID))> Em Portugal a sua codificação sistemática dos episódios de internamento iniciou-se em 1989.

Dentro do conjunto dos códigos aceites num determinado momento, são feitas validações em relação à idade, ao sexo, e à posição de cada código específico dentro da hierarquia da categoria a que pertence, de modo que não sejam recolhidos códigos que têm subcategorias ou subclassificações, e se obtenha o máximo de especificação possível.

**MeSH**

MeSH é um vocabulário controlado de termos organizado hierarquicamente em 16 árvores. A cada árvore é atribuída uma letra de forma a identificar a mesma.



Figura - Estrutura hierárquica do MeSH [1]

Cada termo está localizado numa ou várias árvores simulando, desta forma, uma relação entre diferentes hierarquias, como o SNOMED. A estrutura da árvore tende de conceitos gerais para conceitos mais específicos.

Com a ajuda destas hierarquias pretende-se que o resultado de uma pesquisa por parte do utilizador venha associado a um maior valor no contexto clínico.

## EHR

O EHR como o próprio no indica é uma norma usada para registo eletrónico de documentos clínicos. Surgiu em meados dos anos 60 e 70 essencialmente devido a estudos realizados pelos centros médicos académicos e organizações clínicas governamentais. A Universidade de *Utah Together* juntamente com a Corporação *3M* desenvolveram um dos primeiros EHR (*Health Evaluation through Logical Processing - HELP*). Ao longo dos anos foram desenvolvidas várias outras ferramentas nesta área como: *AllScripts*, *EHRserver*, etc. O EHR surge da necessidade de resolver vários problemas como: a constante necessidade de adaptação das bases de dados, a dificuldade em manter compatibilidade com dados anteriores, a dificuldade em documentar alterações, a heterogeneidade na forma de introdução de dados, as dúvidas na interpretação na leitura, o fato dos conceitos evoluírem e a falta de avaliação na qualidade dos dados. Além destes problemas as estruturas de dados (e a qualidade) dependem muito do objetivo na sua escolha (prestação de cuidados, investigação clinica, gestão/financeira, ensino, legal).

Perante os problemas referidos e, uma vez que o EHR pretende corrigi-los, é possível descreve-lo como uma especificação (norma) que permite a representação de conceitos clínicos complexos.

A norma EHR está construída da seguinte forma (bibliografia-ricardocorreia):

• **Arquétipo**: conceito mais pequeno nesta estrutura (equivalente a uma peça de LEGO), consiste em dados médicos, como por exemplo, altura, peso, sumário de gravidez e ecocardiograma;

• **Template**: consiste num conjunto de peças LEGO, isto é, num conjunto de arquétipos.

Perante estes conceitos é possível construir um esquema que define a estrutura de um documento podendo, assim, adaptar a informação presente nos documentos existentes de forma a homogeneizar os seus dados.

|  |  |
| --- | --- |
| EHR | Registo clínico eletrónico por pessoa |
| Folders | Organização de alto nível do EHR (ex.: por episódio, por especialidade clínica) |
| Compositions | Conjunto de entradas submetidas numa determinada data (ex.: relatório) |
| Sections | Cabeçalhos clínicos que refletem o fluxo de trabalho |
| Entries | Declarações clínicas sobre observações, avaliações e instruções |
| Clusters | Entradas compostas |
| Elements | Entradas elementares |
| Data values | Termos codificados do conjunto de termos |

Tabela - Modelo de referência, bibliografia

Uma das principais vantagens na utilização desta norma é o fato de os utilizadores poderem alterar os arquétipos e *templates* sem alterar o *software*, ou seja, estamos perante uma ferramenta bastante geral, que aceita qualquer tipo de documento (desde que seja elaborado um bom *template* para receber os seus dados).

Neste contexto surge o *openEHR*, que é uma comunidade virtual que trabalha no sentido de tornar os documentos físicos em documentos eletrónicos seguindo a norma EHR, criando arquétipos e *templates*. Este tem definidos uma séria de padrões que seguem a especificação do EHR. Assim sendo, na imagem seguinte estão presentes os atores envolvidos no *openEHR*:



Figura - Atores envolvidos no *openEHR*, bibliografia

A partir desta norma surgem diversas ferramentas que fazem uso dela. Duas delas são apresentadas nos tópicos seguintes.

### AllScripts

Something

### EHRserver (cabolabs)

O *EHRServer* é um repositório de informação clínica gratuito que segue uma arquitetura orientada a serviços (REST API) e é compatível com o padrão do *openEHR*. (bibliografia, pablo pazos) Com a REST API, o *EHRServer* possibilita o armazenamento e a pesquisa de dados clínicos de muitas formas, suportando diversos formatos como XML e JSON. É ainda possível criar *queries* de forma simples através da interface do utilizador, com o *EHRServer Query Builder*. Além disto, como é *opensource* é possível adapta-lo às diferentes necessidades.

No anexo x é apresentado o modelo EER da base de dados do EHRServer em mysql.

Como pode ser observado e, analisando apenas as tabelas diretamente relacionados com o EHR, cada pessoa pode ter vários **ehr**’s e para cada um destes existem várias **contribution**, isto é, cada **ehr** tem vários registos/contribuições clínicas de um doente. Para cada uma destas contribuições é fácil entender a tabela **version**, faz o registo de versões de cada registo. Ao lado da tabela **version** tem a tabela **audit\_details** que é responsável pelo armazenamento de dados como data em que foi submetido um registo, médico que o submeteu, etc. Seguindo em direção a **composition\_index**, é possível observar que várias versões seguem apenas uma composição (**composition\_index**), ou seja, apesar das várias versões nas quais pode ser alterado o conteúdo de um registo, a sua composição/estrutura são sempre as mesmas (segue a estrutura do *template*). Sendo assim cada composição pode conter vários valores, os quais são registados em **data\_value\_index**. Para cada um destes é feita a separação pelas tabelas respetivas de acordo com o tipo de dados que representam.

Com isto, e como já referido anteriormente, é de fácil reparo que o *EHRServer* permite uma uniformização dos registos clínicos, aproveitando-se da estrutura em *templates* e arquétipos para guardar dados com diferentes estruturas.

## Resumo

Em suma, este capítulo permitiu realizar uma análise da recuperação, indexação e classificação de informação de um modo geral e direcionada para o meio clínico.

Relativamente à indexação, é possível afirmar que durante os últimos anos têm vindo a ser desenvolvidas diversas ferramentas que permitem a realização da mesma de forma eficaz. Exemplos disso são o *Apache Lucene*, *Apache Solr*, *Elasticsearch* e *Sphinx Search*. Com estas torna-se possível a indexação de documentos provenientes de diversas fontes e com diferentes formatos e, como são implementadas com a ajuda de alguns dos modelos referidos neste capítulo, os modelos clássicos e estruturados, permitem que a informação apresentada ao utilizador seja a mais relevante perante a pesquisa feita pelo mesmo. No entanto, os modelos referidos apenas dizem respeito a uma indexação geral, isto é, a uma indexação sem conteúdo clínico e, por isso, com este tipo de abordagem podem ser rejeitados resultados com bastante importância. Face a este problema alguns autores sugerem a utilização de métodos de classificação como forma de apresentar resultados com maior relevância para o utilizador. No âmbito dos métodos de classificação surgem assim nomenclaturas que, de uma forma hierárquica, permitem relacionar termos clínicos, como por exemplo: SNOMED, ICD-9 e MeSH.

Acrescentando ao referido, surge ainda o EHR que ajuda no processo de indexação na medida em que junta num só local toda a informação documental que inicialmente se encontrava dispersa pelas diversas fontes. Apesar de ser possível fazê-lo sem o EHR, este último garante a integridade de toda a informação clínica assim como a sua uniformização e estruturação.

Em suma, com o EHR, com as ferramentas de indexação e com as nomenclaturas em causa, é possível criar uma solução com resultados mais úteis e mais direcionados para as necessidades do utilizador (médicos), tendo em conta que a relevância dos resultados e a apresentação dos mesmos ao utilizador são essenciais para esse objetivo.

Capítulo 3

# Solução

Como analisado em outros tópicos deste documento, é possível concluir vários problemas que advêm da grande quantidade de informação clínica. Entre estes podem referir-se a grande variedade de dados, os diferentes formatos desses dados, as diferenças na estrutura dos dados, etc. A partir disto e, uma vez escolhida como ferramenta de indexação, o *Sorl*, e como solução para uniformizar a informação, o *EHRServer*, a arquitetura apresentada a seguir foi construída a pensar não só nestes dois fatores mas também nas necessidades exigidas pela aplicação de forma a apresentar ao utilizador resultados interessantes do ponto de vista clínico.

Sendo assim, neste capítulo pretende-se apresentar a arquitetura da aplicação e evidenciar como é que esta contribui para a resolução do problema em causa.

## Arquitetura

A arquitetura construída descreve uma aplicação cliente-servidor na qual todos os clientes podem conectar-se ao servidor e realizar ações, obtendo a resposta do servidor.

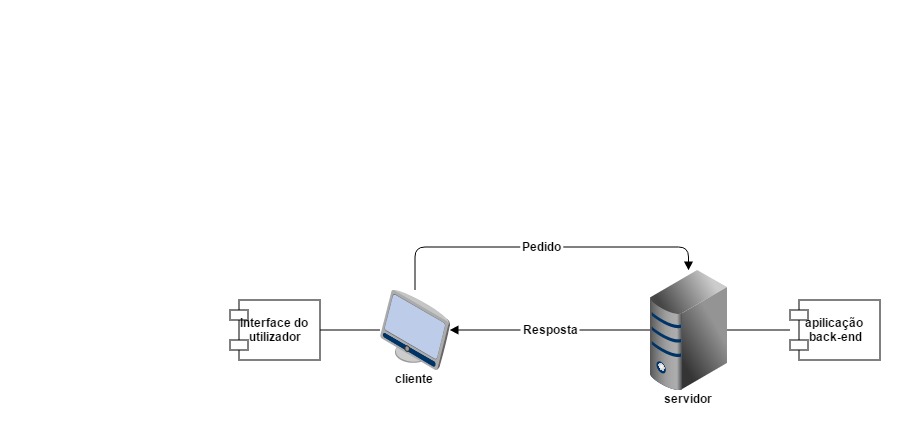


Figura - Estrutura da aplicação cliente/servidor

O lado do cliente é meramente composto pela interface, a qual é usada pelo utilizador para realizar ações e ver resultados. A componente principal da aplicação encontra-se do lado do servidor. Todavia o lado do cliente não deve ser negligenciado, pois sem este, não existem *queries* para o lado do servidor funcionar de acordo com a sua implementação.

Na figura 9 é apresentada a arquitetura da aplicação que é seguida pela explicação mais detalhada de cada um dos passos.

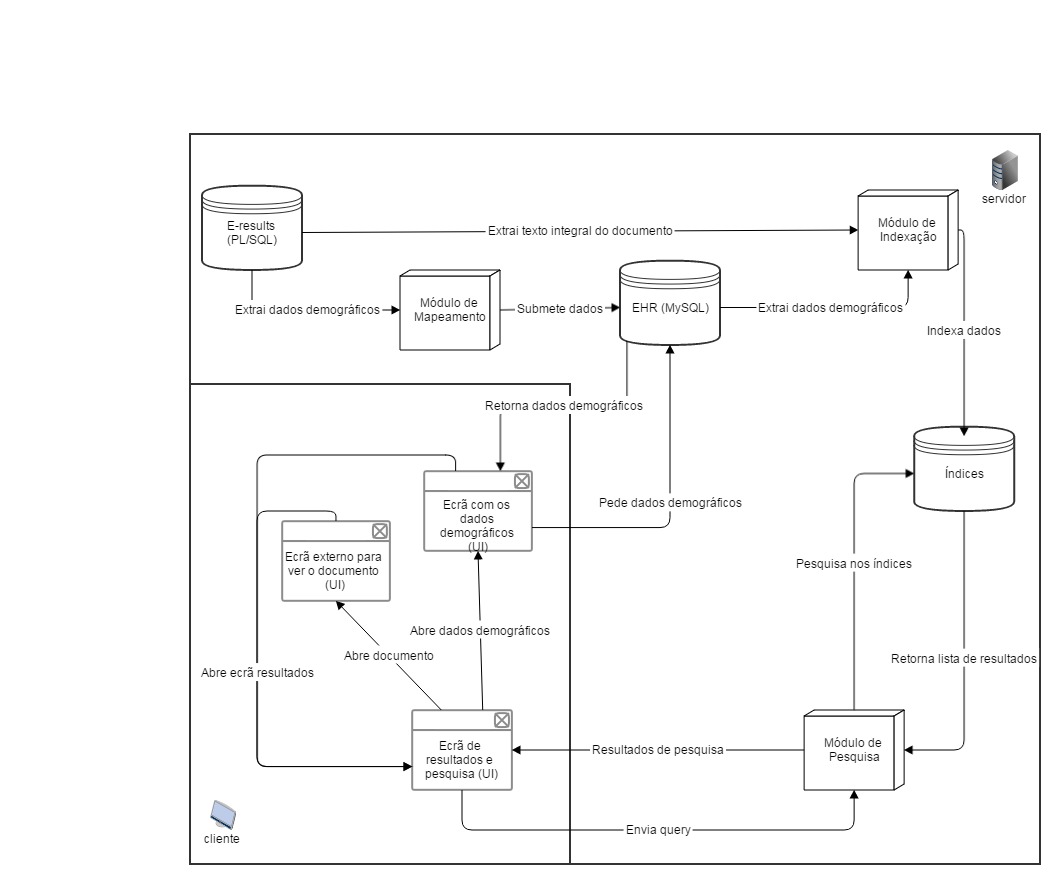


Figura - Arquitetura da solução

### E-results (PL/SQL)

E-results é o nome da base de dados que contem todos os dados que se querem ver tratados. Aqui estão armazenados todos os documentos, desde JSON, XML, PDF, PNG, entre outros. Para cada um destes está associada a respetiva informação demográfica do paciente.

No processo de recolha de informação da base de dados em causa é retornada uma linha por documento, ou seja, para cada documento vêm associados os dados demográficos do doente correspondente.

### Módulo de Mapeamento

Este módulo é responsável pelo mapeamento entre os dados do E-results com as tabelas do EHR. Usa os arquétipos e *templates* para mapear corretamente os dados.

### EHR (MySQL)

Como já referido na revisão bibliográfica, *EHRServer* é um repositório de informação clínica que segue os padrões do *openEHR*. A base de dados EHR é responsável por armazenar os dados relativos ao *EHRServer*, o qual tem um papel fundamental no que diz respeito à uniformização dos dados, sendo, por isso, um dos focos principais desta solução.

O quantidade de dados no EHR varia de acordo com as atualizações e/ou inserções feitas na base de dados do E-results.

### Módulo de Indexação

Previamente, antes de se iniciar a indexação, são definidos, num esquema (*schema*), os parâmetros de indexação e de recolha da informação a ser indexada. Uma vez concluída a construção deste esquema, com a ajuda do *Solr*, procedesse à indexação propriamente dita. Aqui são indexados tanto dados do E-results, como do *EHRServer*. Do E-results são apenas indexados o texto integral dos documentos e os respetivos identificadores (ids) para no futuro ser possível mostrar o documento ao utilizador. Já do *EHRServer* são indexados os dados demográficos dos pacientes.

### Índices

Os índices serão o alvo de todas as *queries* feitas pelo utilizador e esta base de dados é composta por diversos índices. O número destes varia de acordo com as atualizações que vão sendo feitas ao longo do tempo.

### Módulo de Pesquisa

Este módulo é responsável por receber *queries* do utilizador na interface do lado do cliente da aplicação e procurar pelos ficheiros previamente indexados. Após encontrar uma resposta retorna-a para o lado do cliente novamente.

### Interface

A solução em causa apresenta dois tipos de interface: interna, pertence à aplicação, e externa, vem de um serviço externo à aplicação. As interfaces são as seguintes:

* **Ecrã de resultados e pesquisa**: interface interna responsável pela pesquisa e mostragem de resultados. Permite filtrar os resultados de pesquisa por data de nascimento;
* **Ecrã com os dados demográficos**: interface interna responsável por mostrar os dados demográficos de um paciente;
* **Ecrã externo para ver documento**: interface externa que permite a visualização dos documentos indexados.

## Diagrama de casos de uso

Um diagrama de casos de uso descreve o que um utilizador pode fazer num determinado ambiente, neste caso, na aplicação em causa. Neste sentido existe apenas um utilizador, que é o utilizador comum (médicos). Estes podem realizar diversas ações como é possível observar pelo seguinte diagrama.



Figura - Diagrama de casos de uso do utilizador

Os utilizadores podem realizar pesquisa sobre os índices previamente criados pelo indexador. Se ainda não existirem índices então o utilizador, apesar de poder efetuar pesquisa, não receberá nenhum resultado.

## Conclusões

Após a análise da arquitetura da aplicação é possível assumir que com esta os problemas de acesso e outros no que diz respeito à informação podem ser resolvidos. Mas como? De acordo com a figura da arquitetura, é fácil reparar na existência de três módulos. Estes são responsáveis pela resolução dos problemas em causa.

No que diz respeito a documentos com diferentes formatos e estruturas, o módulo do mapeamento resolve isto mapeando os dados para o EHR. Desta forma não interessa quais os dados que estão guardados pois seguem todos a mesma estrutura, a estrutura do *template*.

Relativamente ao facto de nem toda a informação ser necessária ou importante de indexar, é um problema possível de resolver pelo módulo de indexação. Neste, e como já referido, é definido um esquema com os campos a serem indexados assim como os parâmetros para a indexação. Definindo os campos que são indexados estão-se a restringir os dados que vão ocupar lugar na base de dados dos índices. Além disto é também possível definir campos que, não sendo indexados, e por isso não são pesquisáveis, podem ser acedidos, ou seja, mesmo se for necessário manter alguns campos, como por exemplo ids, estes não precisam de ser indexados, são apenas armazenados para futuras consultas.

O restante módulo, o de pesquisa, é responsável por solucionar a apresentação dos documentos com diferentes formatos, isto é, este retorna todos os dados necessários para visualizar os documentos, os dados demográficos e cada linha de resultados. No caso da visualização dos documentos, para poder mostrar ao utilizador documentos com diferentes formatos, recorre-se a um serviço da Glintt que, após receber alguns campos (enviados pelo módulo de pesquisa) pode facilmente mostrar o documento.

Além destes problemas existem outros casos mais específicos que, para serem solucionados, recorrem a mais do que um módulo. Exemplo destes são documentos com apenas imagens. Nestes casos não existe informação textual indexável, sendo necessário associar alguns dados que possam fornecer informação suficiente para que o documento seja pesquisável. Para tal recorre-se aos módulos de mapeamento e indexação, pois ao mapear os dados demográficos de um paciente, uma vez que estão associados ao documento, ao indexa-los é possível realizar uma pesquisa, neste caso por informação demográfica, que retornará não só essa informação como também a informação documental associada.

Em suma, e segundo a análise anterior, é possível afirmar que a arquitetura em causa é viável para que se possam atingir os objetivos desta dissertação e, com isto, resolver os problemas associados à indexação e pesquisa de documentos clínicos.

Capítulo 4

# Implementação

Após definida a arquitetura é importante referir como foi desenvolvida a aplicação e quais as decisões tomadas e o porquê das mesmas ao longo deste processo de desenvolvimento. Com isto, neste capítulo pretende-se especificar cada um dos pontos referidos e descrever detalhadamente todos os processos em causa.

## Protótipo

Antes de iniciar propriamente a descrição da solução, é importante referir que esta é apenas um protótipo, uma prova de conceito para testar e garantir que as funcionalidades em causa serão uteis para o melhor acesso e pesquisa da informação clínica. Assim sendo a aplicação em causa é capaz de indexar e pesquisar documentos que se encontram na base de dados da Glintt.

## Bases de dados da aplicação

### E-results

#### Oracle PL/SQL

#### Modelo de dados

### EHR

#### MySQL

#### Modelo de dados

## Mapeamento

Este é sem dúvida um dos aspetos fundamentais do projeto e também o que levou mais tempo a concluir. Aqui é realizado o mapeamento entre os dados da base de dados da Glintt e o EHR.

É possível dividir este tópico em dois pontos: a base de dados da Glintt e o *EHRServer*. Iniciando pelo segundo, o *EHRServer* é uma aplicação desenvolvida em grails (*groovy and grails*) e que segue a norma EHR. Logo no início surgiram diversas dificuldades associadas a este, desde problemas a executar o servidor, problemas com o funcionamento da base de dados entre outros. No entanto, a dificuldade que mais se evidenciou foi o não saber como é que se iriam passar os dados para o EHR seguindo os *templates*. Após algum estudo e pesquisa sobre o assunto concluiu-se que a forma mais eficaz de o fazer era utilizando a REST API que o *EHRServer* disponibilizava. Todavia este problema ainda nem perto estava de ser resolvido, pois para submeter estes dados era necessário criar, não só arquétipos e *templates*, como também um xml com uma estrutura bem específica que não estava descrita na documentação do *EHRServer*. Para que tal se torna-se exequível foi necessário entrar em contacto com o desenvolvedor do *EHRServer*, Pablo Pazos. Após várias discussões foi possível concluir como se iriam criar os arquétipos, *templates* e xml, e como submete-los no *EHRServer*.

Assim sendo, para construir os arquétipos recorreu-se ao uso do *Archetype Editor* e para os *templates* o *Template Designer*. Já o xml resultou de uma discussão constante com Pablo Pazos. Para testar os pedidos *EHRServer* utilizou-se uma extensão do chrome designada por Insomnia. Com isto finalizado e pronto para testar com vários dados começou-se a pensar quais os dados que seriam importantes ver indexados e, desses, quais os que necessariamente deveriam estar presentes no EHR. Rapidamente se concluiu que o texto integral dos documentos não deveria ser passado para as tabelas do EHR, pois além do processo ser extremamente demorado, teriam que ser criados diversos *templates* para cada estrutura diferente. Além disto, a ideia desta solução é, no futuro, ser aplicada não só na gestão documental como também noutras bases de dados em que o texto se encontra distribuído pelas colunas das diferentes tabelas. Sendo assim ainda menos sentido faz a passagem dos documentos para o EHR. No meio deste processo de escolhas e decisões chegou-se a uma questão fundamental no projeto: será que é mesmo importante e necessária a utilização do EHR? A resposta é sim, pois além da uniformizar a informação permite que esta esteja distribuída por doentes que é o que se pretende encontrar após uma pesquisa realizada com a aplicação em causa. A acrescentar a esta justificação tem-se o facto referido em cima de que a aplicação é, no futuro, para ser aplicada a outras bases de dados que poderão mostrar-se muito mais recetivas ao uso do EHR. Assim sendo, e seguindo com o EHR, chega-se ao segundo ponto, a base de dados da Glintt.

Relativamente à base de dados da Glintt, foi gerado o modelo EER para se perceber melhor como é que esta estava estruturada. Após a análise da mesma voltou-se novamente à mesma questão no que diz respeito aos dados que devem ser indexados. Como referido, o texto integral não tem interesse em ser indexado, mas então como é que se vai pesquisar informação? A resposta a esta pergunta está nos dados demográficos dos pacientes que, ao estarem associados por ids aos respetivos documentos, permitem que estes possam ser mostrados na pesquisa. Assim sendo, conclui-se que a única informação necessária para armazenar no EHR são os dados demográficos.

Em suma, a aplicação funciona com dois tipos de bases de dados, o EHR, que armazena os dados demográficos dos pacientes recorrendo a arquétipos, *templates* e xml, e a base de dados da Glintt, que armazena os documentos.

Na imagem seguinte encontra-se o *template* usado na aplicação.

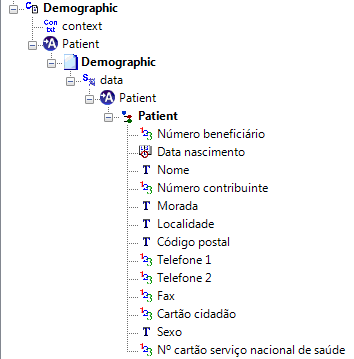


Figura - *Template* (dados demográficos)

Como se pode observar pela figura anterior este *template* é apenas formado por um arquétipo, o arquétipo **Patient**. Além disso é fácil perceber que tipo de dados demográficos estão a ser armazenados no EHR.

## Indexação

Nesta etapa existiram muito menos dificuldades, essencialmente devido à boa documentação. Para a indexação recorreu-se ao uso do *Solr*, o qual permite através da edição de alguns ficheiros xml definir todos os parâmetros e definições necessárias para a indexação dos dados em causa.

Existem três ficheiros de configuração importantes no *Solr*:

* **data-config.xml** – responsável pelas *queries* essenciais para a recolha de informação das diferentes fontes de dados (neste caso a base de dados da Glintt e o EHR);
* **solr-config.xml** – responsável pelas configurações do *Solr*. É neste ficheiro que se estão definidos todos os parâmetros ligados à pesquisa, como por exemplo, em que campo ou campos é que o *Solr* vai realizar pesquisa, qual o tamanho do texto que vai aparecer no ecrã de resultados, etc;
* **schema.xml** – responsável pela estrutura dos dados que vão ser indexados, armazenados ou não no Solr. É o ficheiro de configuração mais importante, pois é aqui que também se definem como será efetuado o processo de indexar.

Inicialmente começou-se por testar estes ficheiros xml com as configurações predefinidas. Seguidamente começaram-se a fazer testes com o **data-config.xml**, o qual, após algum tempo já estava preparado para receber dados do EHR. Após este ponto, surgiram alguns problemas no que diz respeito à quantidade de pedidos e velocidade de indexação. Demorava entre 8 e 12 segundos para indexar 8 documentos. Além disto, eram feitos cerca de 467 pedidos à base de dados para a mesma quantidade de documentos (nesta altura ainda só se estavam a indexar/armazenar dados do EHR). Com isto, veio associada uma grande preocupação em reduzir a quantidade de pedidos e aumentar a velocidade de indexação. Para atingir este objetivo, além de haver uma correção nas *queries*, foi realizada uma nova análise dos dados que tinham interesse em ser indexados, ficando decidido que inicialmente apenas se indexaria texto (mais tarde foram inseridas datas para uma filtragem dos resultados). Após isto a velocidade de execução tornou-se muito maior, cerca de 2 segundos para os mesmos 8 documentos. Já para não falar na quantidade de pedidos que foi reduzida para aproximadamente 86.

À medida que iam sendo feitas alterações no **data-config.xml**, os outros dois ficheiros iam também sendo modificados de acordo com as necessidades. Assim, e como o **schema.xml** é o mais importante, segue-se um pequeno excerto do mesmo, mostrando como os dados no Solr estaão estruturados.

<field name=**"uid"** type=**"string"** indexed=**"true"** stored=**"true"**/>

<field name=**"value"** type=**"text\_general"** indexed=**"true"** stored=**"true"** multiValued=**"true"**/>

<field name=**"dates"** type=**"date"** indexed=**"true"** stored=**"true"** multiValued=**"true"**/>

<field name=**"first\_name"** type=**"string"** indexed=**"false"** stored=**"true"**/>

<field name=**"last\_name"** type=**"string"** indexed=**"false"** stored=**"true"**/>

<field name=**"dob"** type=**"date"** indexed=**"false"** stored=**"true"**/>

<field name=**"elemento\_id"** type=**"int"** indexed=**"true"** stored=**"true"** />

<field name=**"documento\_id"** type=**"int"** indexed=**"true"** stored=**"true"** />

<field name=**"content"** type=**"text\_general"** indexed=**"true"** stored=**"true"**/>

<field name=**"entidade\_id"** type=**"string"** indexed=**"false"** stored=**"true"**/>

<field name=**"doente"** type=**"string"** indexed=**"false"** stored=**"true"**/>

<field name=**"text"** type=**"text\_general"** indexed=**"true"** stored=**"false"** multiValued=**"true"** />

Como é possível observar nem todos os campos têm a opção *indexed* e *stored* como verdadeira, ou seja, nem para todos os campos é necessário armazenar e indexar. No caso dos campos que não são indexados, não podem ser pesquisados, no caso dos campos que não são armazenados, não podem ser vistos no resultado de pesquisa. Face a isto, podem-se tirar diversas vantagens na utilização destes dois campos, por exemplo, os campos **entidade\_id** e **doente** (também é um id que juntamente com a **entidade\_id** formam a chave primária do paciente) não estão indexados, mas estão armazenados, ou seja, não tem interesse indexar ids, no entanto, manter a informação destes é extremamente importante para identificar, neste caso, o doente. Relativamente aos campos **elemento\_id** e **documento\_id** será justificada a necessidade de os indexar no tópico seguinte, assim como os campos **first\_name**, **last\_name** e **dob**, por não serem indexados.

Os campos relativos ao texto integral do documento e à informação demográfica são **content** e **value** respetivamente. Estes são copiados, noutra parte do **schema.xml**, para o campo **text**, o qual não fica armazenado, pois apenas é utilizado para fins de pesquisa e também não interessa armazenar duas vezes a mesma informação. É por isso definido como campo *default* para a pesquisa.

Falando agora do campo **text**, é possível observar que o seu tipo é **text\_general**, ou seja, corresponde a um campo de texto no qual, na sua definição, estão presentes todos os dados relativos aos parâmetros de indexação e pesquisa. Na seguinte tabela faz-se uma análise mais cuidada de cada um desses parâmetros.

|  |
| --- |
| <tokenizer class="solr.StandardTokenizerFactory"/> |
| Divide o texto em *tokens*, ou seja, em palavras, tratando os espaços em branco e a pontuação como delimitadores. Esses delimitadores são apagados. |
| Exemplo (retirado de <https://cwiki.apache.org/confluence/display/solr/Tokenizers>):  *In*: "Please, email john.doe@foo.com by 03-09, re: m37-xq."  *Out*: "Please", "email", "john.doe", "foo.com", "by", "03", "09", "re", "m37", "xq" |
| <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt" /> |
| Como o próprio nome indica, é o parâmetro responsável pela remoção das *stopwords*, as quais estão definidas no ficheiro stopwords.txt (proveniente do *Solr*). |
| <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" ignoreCase="true" expand="true"/> |
| Encontra *tokens* e modifica-as com outras, ou seja, permite que uma pesquisa por “dor de barriga” encontre também resultados com “dor abdominal”. |
| <filter class="solr.LowerCaseFilterFactory"/> |
| Coloca as letras de cada *token* em minúscula. |
| <filter class="solr.RemoveDuplicatesTokenFilterFactory" /> |
| Remove duplicados para não sobrecarregar o processo de indexação com informação repetida. |
| stemming |

Tabela - Parâmetros de indexação do *Solr*

Ainda na análise dos parâmetros, é possível referir que nem todos estão presentes no processo de pesquisa e indexação, exemplo disto é o parâmetro relativo aos sinónimos que só está definido para a pesquisa e não para a indexação, uma vez que não faz sentido indexar mais palavras que retornarão o mesmo resultado.

Saindo da área dos parâmetros de pesquisa e indexação, para resultar num melhor desempenho desta etapa da indexação pensou-se ainda em paralelizar este processo. Para tal, e com o fim de criar uma aplicação cliente-servidor, recorreu-se ao .NET, que consiste numa *framework* em c#. Com isto foi possível utilizar o *Solrnet* que consiste num cliente do *Solr* que permite utilizar as funcionalidades do mesmo através de funções em c# (paralelizáveis). Sendo assim, a utilização do ficheiro data-config.xml tornou-se inútil, uma vez que as *queries* podiam ser feitas através do c#.

Em suma, o processo de indexação é bastante complexo, no entanto a utilização de ferramentas como o *Solrnet* facilitam a sua implementação tornando possível a gestão de processos paralelos e, assim, melhorando o desempenho da aplicação.

## Pesquisa

sfddf

## Interface

Fdsf.

## Resumo

### Problemas

### Software

### Conclusão

Capítulo 5

# Simulação e Resultados

Neste capítulo serão apresentadas algumas conclusões sobre os objetivos já concluídos e também dos não concluídos de forma a avaliar o trabalho que será desenvolvido num futuro próximo.

## Cenário

Como já foi referido em vários pontos deste documento, o objetivo desta dissertação é criar uma solução que disponibilize às entidades médicas uma nova forma de acesso à informação, mais rápida e eficaz do que a existente atualmente.

## Resultados Experimentais

fdg

Capítulo 6

# Conclusões e Trabalho Futuro

Neste capítulo serão apresentadas algumas conclusões sobre os objetivos já concluídos e também dos não concluídos de forma a avaliar o trabalho que será desenvolvido num futuro próximo.

## Satisfação dos Objetivos

Como já foi referido em vários pontos deste documento, o objetivo desta dissertação é criar uma solução que disponibilize às entidades médicas uma nova forma de acesso à informação, mais rápida e eficaz do que a existente atualmente.

Para que os objetivos referidos sejam atingidos, nesta primeira parte do trabalho, foram estudadas diversas ferramentas, modelos e métodos que de alguma forma podem contribuir para a concretização dos mesmos.

Face ao estudo efetuado, só posso estar satisfeito com o trabalho desenvolvido, pois este permitiu-me conhecer melhor como funcionam os processos de recuperação, indexação e classificação de informação, assim como adaptar cada um deles ao contexto em causa.

## Trabalho Futuro

Como ainda só foi realizada a primeira parte do trabalho, existem imensas tarefas por realizar e, por isso, na figura seguinte é apresentado o plano de trabalho que penso realizar na segunda parte deste projeto.

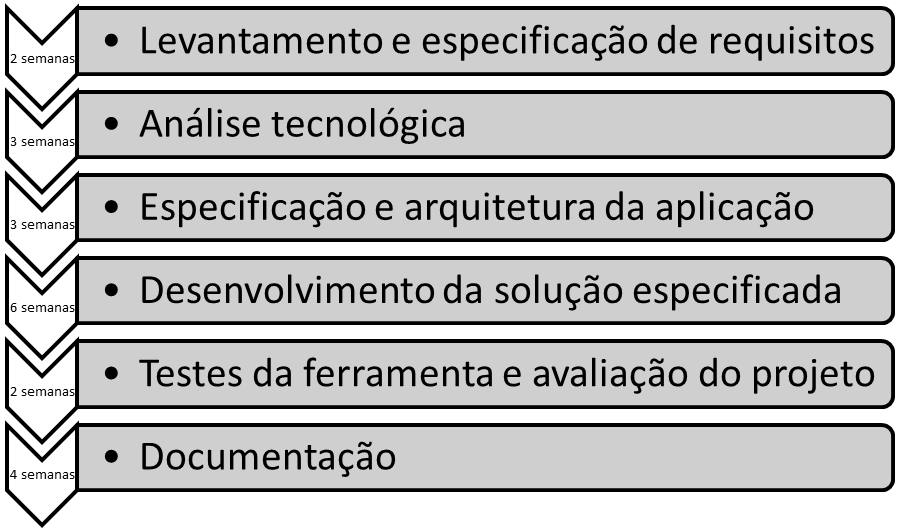


Figura - Planeamento

Relativamente aos dois primeiros tópicos do planeamento, já foram iniciados ao longo deste estudo do estado da arte e, por isso, possivelmente serão cumpridos antes da data prevista. No entanto e, especialmente no que diz respeito à análise tecnológica, é ainda necessário realizar alguns testes práticos com as ferramentas de forma a concluir a analisa iniciada e descrita neste documento. Após efetuar a análise tecnológica, procede-se à especificação e arquitetura da aplicação, seguida do desenvolvimento da mesma.

Uma vez concluída a etapa anterior, pretendem-se realizar testes ao *software* com o objetivo de garantir a estabilidade da aplicação. Nas mesmas duas semanas em que serão realizados os testes pretende-se ainda que seja feita uma avaliação ao projeto, ou seja, uma avaliação que permita perceber se a solução produzida realmente causa um impacto positivo no meio clinico. Para este ponto, é ainda necessário decidir como será feita essa avaliação.

Por fim, nas últimas quatro semanas, espera-se que haja a escrita da documentação referente à solução proposta assim como a continuação da escrita deste documento

# Referências

[1] (2016). *Home - MeSH - NCBI*. Available: <http://www.ncbi.nlm.nih.gov/pubmed/>

[2] (2016). *Sphinx | Open Source Search Engine*. Available: <http://sphinxsearch.com/>

[3] (2016). *DB-Engines Ranking*. Available: <http://db-engines.com/en/ranking/search+engine>

[4] E. Hatcher and O. Gospodnetic, *Lucene in Action (In Action series)*: Manning Publications Co., 2004.

[5] Ihtsdo. (2016). *SNOMED CT Document Library*. Available: <http://ihtsdo.org/fileadmin/user_upload/doc/>

[6] C. R. L. Coelho, "Acesso e Recuperação de Informação em Catálogos Bibliográficos Online," Faculdade de Engenharia da Universidade do Porto, 2014.

[7] M. Lesk, "The seven ages of information retrieval," ed, 1996.

[8] C. D. Manning, P. Raghavan, H. Sch\, \#252, and tze, *Introduction to Information Retrieval*: Cambridge University Press, 2008.

[9] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*: Addison-Wesley Longman Publishing Co., Inc., 1999.

[10] D. Hiemstra, *Using language models for information retrieval*: Taaluitgeverij Neslia Paniculata, 2001.

[11] D. L. Lee, H. Chuang, and K. Seamons, "Document ranking and the vector-space model," *Software, IEEE,* vol. 14, pp. 67-75, 1997.

[12] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *Journal of the American Society for Information science,* vol. 27, pp. 129-146, 1976.

[13] R. Baeza‐Yates and G. Navarro, "XQL and proximal nodes," *Journal of the American Society for Information Science and Technology,* vol. 53, pp. 504-514, 2002.

[14] D. A. Hull, "Stemming algorithms: A case study for detailed evaluation," *JASIS,* vol. 47, pp. 70-84, 1996.

[15] M. McCandless, E. Hatcher, and O. Gospodnetic, *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*: Manning Publications Co., 2010.

[16] T. Grainger and T. Potter, *Solr in action*: Manning Publications Co., 2014.

[17] C. Mattmann and J. Zitting, *Tika in Action*: Manning Publications Co., 2011.

[18] (2016). *Apache Solr*. Available: <http://lucene.apache.org/solr/>

[19] R. Gheorghe, M. L. Hinman, and R. Russo, *Elasticsearch in Action*: Manning Publications Co., 2015.

[20] D. Rothwell, F. Wingert, R. Cote, R. Beckett, and J. Palotay, "Indexing Medical Information: The Role of SNOMED," in *Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, 1989, pp. 534-539.

[21] (2016). *ICD - ICD-9-CM - International Classification of Diseases, Ninth Revision, Clinical Modification*. Available: <http://www.cdc.gov/nchs/icd/icd9cm.htm>

Anexos

# Anexo A

Depois das conclusões e antes das referências bibliográficas, apresenta-se neste anexo numerado o texto usado para preencher a dissertação