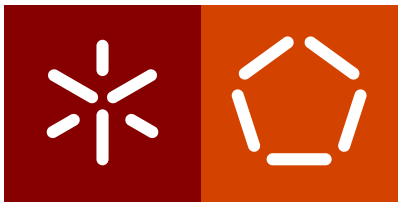
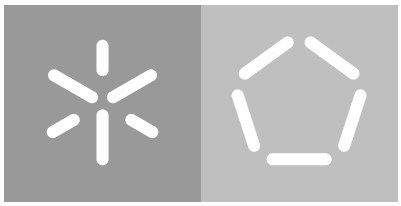


Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Pedro Carvalho Gomes

Georreferenciação de Conteúdos de Bases de Dados Documentais

15 de Janeiro de 2020



Universidade do Minho
Escola de Engenharia
Departamento de Informática

João Pedro Carvalho Gomes

Georreferenciação de Conteúdos de Bases de Dados Documentais

Dissertação
Mestrado Integrado em Engenharia Informática

Dissertação supervisionada por
Professor Doutor Orlando Belo
Professora Doutora Anabela Barros

15 de Janeiro de 2020

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

AGRADECIMENTOS

Em primeiro lugar, quero agradecer ao Departamento de Informática e à Universidade, pois, sem todo o conhecimento e experiência que me proporcionaram ao longo do Mestrado Integrado em Engenharia Informática, não seria possível iniciar um projeto ambicioso e de contexto real.

Aos meus orientadores Professor Orlando Manuel Oliveira Belo e Anabela Leal Barros, por toda a dedicação, empenho, prontidão e tempo disponibilizado no auxílio e acompanhamento de todas as fases de desenvolvimento desta dissertação.

À Daniela Gomes, ao acompanhar-me todos os dias desde do início, por me motivar, incentivar e apoiar. Por todo o companheirismo, amor, carinho e por todos os momentos de abstração e diversão que me ajudaram a ultrapassar os momentos menos bons e a terminar esta etapa da minha vida com sucesso.

Ao Diogo Martins, Francisco Martins, Miguel Morais e Rafael Botelho, por me terem proporcionado grandes momentos de diversão e distração quando mais precisava e pelo todo o apoio ao longo da minha vida.

Ao Tiago Fraga, João Reis, Miguel Chaves, Joel Morais e Joel Peixoto por me terem proporcionado momentos incríveis durante a minha vida académica, pelo acompanhamento nos dias de biblioteca e por toda a diversão e apoio ao longo dos dias, ajudando assim a completar mais uma etapa na minha vida.

A Sara Coelho e Helena Ribeiro, por me terem proporcionado bons momentos de diversão, ajudando assim na abstração dos dias menos bons e por todo o apoio ao longo desta fase da minha vida académica.

Aos meus pais e ao meu irmão que desde sempre me apoiaram incondicionalmente e que sempre me proporcionaram todas as condições necessárias ao meu sucesso, tanto nas vertentes profissionais e académicas, como também pessoais.

RESUMO

A georreferenciação é o processo de localização geográfica de um determinado objeto espacial através da atribuição de coordenadas. Os sistemas de georreferenciação utilizam um processamento espacial automático executado por computador, por exemplo para colocar uma entidade num mapa ou fornecer um recurso espacial. Quando este processo é aplicado a coleções de documentos textuais, é descrito como uma combinação de reconhecimento de entidades nomeadas. O Livro das Propriedades, também designado como o *Tombo da Mitra*, contém informação relativa aos tipos de terras, acidentes de terreno, nomes de ruas, proprietários e apontamentos biográficos e genealógicos das várias propriedades que a mesa Arcebispal de Braga possuía no século XVII. Este trabalho de dissertação teve como objetivo conceber e implementar um sistema de georreferenciação textual para o conteúdo existente no Livro das Propriedades, com particular enfoque nos lugares que nele estão referidos, de forma a permitir aos estudiosos destes conteúdos possuírem informação acerca da localização geográfica desses elementos.

Palavras-chave:Aprendizagem Máquina, Mineração de Textos, Processamento de Linguagem Natural (PLN), Sistemas de Georreferenciação, Livro das Propriedades, Tombo da Mitra.

ABSTRACT

Georeferencing is the process of geographically locating a given spatial object by assigning coordinates. Georeferencing systems use automatic spatial processing performed by a computer, for example to place an entity on a map or provide a spatial resource. When this process is applied to collections of textual documents, it is described as a combination of named entity recognition. The Livro das Propriedades, also designated as "Tombo da Mitra", contains information regarding land types, landforms, street names, owners, and biographical and genealogical notes of the several properties that the Archbishop's table of "Braga"owned in the 17th century. This dissertation work aimed to design and implement a textual georeferencing system for the existing contents of the "Livro das Propriedades", with particular focus on the places mentioned in it, in order to allow scholars of these contents to have information about the geographical location of those elements.

Keywords: Machine Learning, Text Mining, Natural Languague Processing (NLP), Georeferencing Systems, "Livro das Propriedades", "Tombo da Mintra".

CONTEÚDO

1 INTRODUÇÃO 12
- 1.1 Contextualização 12
- 1.2 Motivação e Objetivos 13
- 1.3 Organização da Dissertação 14

2 GEORREFERENCIAÇÃO 16
- 2.1 Uma Definição 16
- 2.2 Utilidade Prática 17
- 2.3 Utilização 18
- 2.4 Ferramentas de Georreferenciação 18
  - 2.4.1 ArcGIS 19
  - 2.4.2 OpenCalais 20
  - 2.4.3 CLAVIN 20
  - 2.4.4 The Edinburgh Geoparser 20
  - 2.4.5 Geoparser.io 21
  - 2.4.6 GeoTXT 21
  - 2.4.7 Ferramentas de Processamento de Linguagem Natural 22
- 2.5 Análise Comparativa das Ferramentas 23

3 O SISTEMA DE GEORREFERENCIAÇÃO 25
- 3.1 Geoparsing e Geoconding 25
- 3.2 Modelos e Processos de Georreferenciação 26
- 3.3 O Modelo Desenvolvido 28
- 3.4 Descoberta de Referências de Locais 28
- 3.5 Mapeamento de Referências 29

4 CASO DE ESTUDO 31
- 4.1 Apresentação Geral 31
- 4.2 O Livro das Propriedades 31
- 4.3 Tommi 34
  - 4.3.1 Descrição do Sistema 34
  - 4.3.2 A Integração de Serviços de Georreferenciação 39

5 IMPLEMENTAÇÃO 41
- 5.1 Criação/Processamento do Dicionário de Locais 41
- 5.2 Atualização de Grafia 43
- 5.3 Anotação de Locais 44
  - 5.3.1 Identificação de Potenciais Locais 45
  - 5.3.2 Anotação de Locais com base no dicionário de locais 47

5.4 A integração do módulo de Georreferenciação 48
5.4.1 O sistema de *Backend* 48
5.4.2 O sistema de *FrontEnd* 49
6 CONCLUSÕES E TRABALHO FUTURO 53
6.1 Conclusões 53
6.2 Trabalho Futuro 54

LISTA DE FIGURAS

Figura 1 Um dos ambientes de trabalho da Plataforma ArcGIS ESRI (2020a) 19
Figura 2 Exemplo de geojson 21
Figura 3 Modelo de Sistema de Georreferenciação adaptado de Grover et al. (2010) 26
Figura 4 Abordagem do Modelo de Georreferenciação extraído de Chen et al. (2018) 27
Figura 5 Esquema do Modelo do Sistema de Georreferencição 28
Figura 6 Esquema do Geoparser 29
Figura 7 Exemplo do Ficheiro com os locais Anotados depois de atuar *geoparser*. 29
Figura 8 Esquema do *Geocoder* 30
Figura 9 *Livro de Propriedades fechado e Livro de Propriedades aberto* 33
Figura 10 Excerto do Folio 97 e sua transcrição 33
Figura 11 Ambiente de Autenticação do Sistema 35
Figura 12 Ambiente Principal de Trabalho no Sistema 35
Figura 13 Passo 1 da Importação - Catalogação 36
Figura 14 Passo 2 da Importação - Visualização 37
Figura 15 Passo 4 da Importação - Etiquetas Identificadas 37
Figura 16 Passo 5 da Importação - Visualização de Índices 38
Figura 17 Gestão de Fólios 38
Figura 18 Gestão de Índices 39
Figura 19 Gestão de Utilizadores 39
Figura 20 Diagrama do Processo realizado pelo sistema de georreferenciação de texto 41
Figura 21 O Processo de Criação do Dicionário de Locais 42
Figura 22 O processo de atualização de grafia 43
Figura 23 Exemplos de regras de atualização de grafia 44
Figura 24 O processo de *geoparsing* do sistema de georreferenciação de texto 44
Figura 25 O processo de identificação de Potenciais Locais 45
Figura 26 O processo de pesquisa com base em dicionário de locais 47
Figura 27 Esqueleto de uma View em VueJS 49
Figura 28 Esqueleto do Script de uma View/Component 49
Figura 29 Exemplo de uma rota definida em Vue 50
Figura 30 Fragmento da lista de locais anotados no sistema 50
Figura 31 O processo de Georreferenciação do sistema *Tommi* 51

Figura 32 Mapa com representação de uma localidade, em modo satélite e estrada 51
Figura 33 Opções de mudança de Forma de Mapa 52
Figura 34 Vista de todas as localidades georreferenciadas 52

LISTA DE TABELAS

Tabela 1	Comparação de características das ferramentas de georreferenciação estudadas	23
Tabela 2	Exemplo de alguns campos presentes nos dados do *Geonames*	42
Tabela 3	Dicionário de Conversão	44
Tabela 4	Identificadores de Classe	46
Tabela 5	Critérios de Paragem	46

SIGLAS

API Application Programming Interface. 1, 21, 23

BPMN Business Process Modeling Notation. 1, 26

CSV Comma Separated Values Files. 1, 41, 42

ESRI Environmental Systems Research Institute. 1, 19

geojson Geographic JSON. 1, 21, 23
GIR Geographical Information Retrieval. 1, 17

IA Inteligência Artificial. 1

JSON JavaScript Object Notation. 1, 22, 43, 46

MIEI Mestrado Integrado em Engenharia Informática. 1, 13

NEC Named Entity Classification. 1, 22
NER Named Entity Recognition. 1, 12, 18, 21, 22, 28, 46
NLP Natural Languague Processing. 1, 5

PLN Processamento de Linguagem Natural. 1, 4, 18, 20, 22, 24–27, 53
PoS Part-of-Speech tagging. 1, 22, 24

REST Representational State Transfer. 1, 21, 23

SED Stream Editor. 1
SIG Sistemas de Informação Geográfica. 1, 12, 16, 17, 19, 25

UM Universidade do Minho. 1, 13, 34

XML Extensible Markup Language. 1, 22

1

INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Um sistema de georreferenciação de texto pode ser definido na área de Sistemas de Informação Geográfica (SIG) como um sistema que utiliza técnicas de *geoparsing* (Leidner, 2017). Este termo é usado para descrever o processo de identificar nomes em texto. Na área de informática, este processo é conhecido por reconhecimento e classificação de entidades (Leidner, 2007). Neste tipo de sistema é vulgar encontrarmos também o termo *geocoding*, que consiste na tarefa de associar um conjunto de dados georreferenciados implicitamente para associar a dados georreferenciados explicitamente, ou seja, representações com latitude e longitude, tal como acontece nas associações de elementos a mapas convencionais. Estes sistemas apresentam grande utilidade nos dias de hoje, uma vez que a georreferenciação, com a ajuda da digitalização e com a publicação online, facilita o acesso de utilizadores não especializados a informação geográfica, patrimonial, urbana ou ambiental de interesse. Com esta capacidade podemos incorporar mapas atuais, como *Google maps*, *Bing maps*, bem como mapas antigos e fazer a sua comparação, impulsionando o desenvolvimento de estudos históricos diacrónicos e evolutivos por urbanistas, historiadores ou arquitetos, sobre uma cidade, um estado ou um território, mostrando desta forma a sua evolução ao longo do tempo (Cascón-Katchadourian et al., 2018).

Dentro da área da georreferenciação de texto, existe já um património de *software* bastante interessante com a capacidade de realizar tarefas como as referidas, tal como *OpenCalais* (Refinitiv (2019)), *The Edinburgh Geoparser* (Grover et al. (2010)), *Geoparser.io*, *CLAVIN* (Novetta (2019)) e *GeoTXT* (Karimzadeh et al., 2019). Veja-se,por exemplo, em detalhe, este último. Este programa foi criado com o objetivo de reconhecer e localizar geograficamente nomes de lugares em textos não estruturados. Para realizar esta tarefa o programa utiliza algoritmos de Named Entity Recognition (NER) para reconhecimento de nome e local, e um mecanismo de procura cooperativa para indexar, classificar e recuperar topónimos, conseguindo obter um *geoparsing* escalável em relação ao fluxo de texto.

Para que se entenda o módulo de georreferenciação descrito ao longo desta dissertação é necessário entender o que é o *"Livro de Propriedades"*.

O *Livro das Propriedades* ou *Tombo da Mitra* foi escrito no início do século XVII. Nesta altura as propriedades da Mesa Arcebispal encontravam-se por todo o Minho e Trás-os-montes, che-

gando mesmo ao bispado do Porto e a Santarém, bem como à Galiza (Barros, 2019) (Barros, 2021). Com o objetivo de se registar todas essas propriedades, foi criado então o *Livro das Propriedades*. Este códice tem registados, de uma forma extensa e pormenorizada, as propriedades, as rendas e os foros da Mesa Arcebispal, tanto na zona da cidade de Braga como na região Norte de Portugal, estendendo-se para fora da mesma, tal como explicado anteriormente.

O módulo de georreferenciação de texto vem então permitir realizar as tarefas descritas anteriormente, tais como o processo de *geoparsing* e de *geocoding* sobre o conteúdo do *Livro de Propriedades*. Este módulo apresenta uma grande utilidade para os estudiosos do manuscrito, uma vez que vem permitir uma forma interativa e didática de explorar o conteúdo descrito em cada item e secção do Livro.

Ao longo desta dissertação será possível perceber a grande utilidade que a criação de um módulo de georreferenciação de texto vai ter para os estudiosos do Livro das Propriedades (Barros, 2019) (Barros, 2021). Com esse módulo, eles poderão adquirir conhecimento complementar acerca da área, local e propriedades que estão referidos nos inúmeros registos de propriedades que figuram no livro.

1.2 Motivação e Objetivos

O sistema *Tommi – Sistema de Análise do Livro das Propriedades ou Tombo da Mitra Arquiepiscopal de Braga* (Barros et al., 2020) – foi criado inicialmente por um grupo interdisciplinar composto por quatro elementos, uma professora do Departamento de Estudos Portugueses e Lusófonos, um professor do Departamento de Informática e dois alunos do Mestrado Integrado em Engenharia Informática (MIEI), da Universidade do Minho (UM), com o objetivo de fornecer aos estudiosos do Livro das Propriedades uma ferramenta que os pudesse ajudar a analisar o seu conteúdo, de uma forma simples, permitindo a realização de pesquisas sobre os textos armazenados, sem ter de recorrer repetidamente à leitura dos documentos originais ou se limitar à consulta da sua edição (Barros, 2019) (Barros, 2021). O *Tommi* é uma forma moderna, simples e eficaz de estudar os textos contidos no Livro das Propriedades, que está disponível online (tommi.di.uminho.pt), que permite aceder aos textos do livro, a partir de qualquer lugar, utilizando um browser Web. Antes da edição do códice e da criação deste sistema, a única forma de acesso ao conteúdo do *Livro das Propriedades* era através do manuscrito. Esta possibilidade não permitia a consulta por mais do que uma pessoa ao mesmo tempo, pois trata-se de um códice único, bastante volumoso, nem uma pesquisa expedita dos elementos que nele estão referidos.

De modo a facilitar os processos de análise dos textos contidos no sistema *'Tommi'*, em particular as referências aos locais das propriedades inventariadas, decidiu-se incorporar um sistema de anotação automática, direcionado especialmente para a identificação e georreferenciação desses locais. Assim, para que fosse possível fazer a sua implementação, foi neces-

sário incluir e manter no sistema uma base de anotações de georreferenciação como meio de indexação dos vários elementos de dados contidos no *Livro das Propriedades*.

Tendo todos estes aspetos em consideração, definimos um conjunto específico de objetivos com vista à implementação do referido módulo de georreferenciação de elementos em textos.

- fazer a anotação (semi)automática do *Livro das Propriedades*, de forma a se poder identificar e relacionar os vários locais que nele estão referidos, de forma direta ou indireta;
- associar os locais anotados a um conjunto definido de coordenadas geográficas, bem como associar os locais a um mapa;
- gerir um mapa de relacionamentos entre as anotações encontradas, permitindo a descoberta de conteúdo relacionado;
- por último este módulo deverá ser incorporado no sistema já desenvolvido para o *Tombo da Mitra* (tommi.di.uminho.pt).

1.3 Organização da dissertação

Para além da presente introdução, esta dissertação incorpora mais seis capítulos, que estão organizados da seguinte maneira:

- Capítulo 2 - **Georreferenciação** - que, em linhas gerais, apresenta o domínio da georreferenciação, abordando em particular a georreferenciação de textos, explicando o que é um sistema de georreferenciação de texto e expondo a sua utilidade em várias áreas de aplicação. Além disso, também se explicará como são utilizados os sistemas de georreferenciação de texto e algumas ferramentas que podem ser aplicadas na sua criação e exploração. Por último, far-se-á uma breve análise comparativa dessas ferramentas.
- Capítulo 3 - **O sistema de Georreferenciação** - apresenta exemplos de modelos de georreferenciação analisados e já implementados, bem como o modelo escolhido e o porquê de se ter optado pelo mesmo para a criação do sistema de Georreferenciação de texto.
- Capítulo 4 - **Caso de Estudo** - apresenta ao pormenor o caso de estudo em que se insere esta dissertação, permitindo perceber a necessidade de se criar este tipo de sistema e o contexto em que se insere.
- Capítulo 5 - **Implementação** - que, de forma geral, apresenta todo o trabalho desenvolvido, incluindo a descrição do processo da *criação do dicionário de locais*, o processo de *atualização de grafia* (que inclui as regras para a realização do mesmo), o processo de anotação de locais na sua totalidade e a integração do módulo de Georreferenciação no sistema *Tommi*.

- Capítulo 6 - **Conclusões e Trabalho Futuro** - por último, serão sumarizadas as perspetivas futuras para este projeto, com vista a permitir o desenvolvimento de um sistema melhor e mais capaz. Este capítulo também refletirá sobre o trabalho desenvolvido, os seus pontos positivos, dificuldades encontradas, resultado alcançado, bem como a seleção do modelo utilizado e os comentários relativos ao desenrolar do projeto.

2

GEORREFERENCIAÇÃO

2.1 UMA DEFINIÇÃO

A georreferenciação é o processo de localização geográfica de um determinado objeto espacial através da atribuição de um conjunto de coordenadas. Um recurso georreferenciado é uma quantidade de informação relevante que identifica um local específico na terra. Como tal, qualquer tipo de documento atribuído a um local geográfico específico é um exemplo de um recurso georreferenciado (Solina and Ravnik, 2010). Na georreferenciação de textos, os dados georreferenciados são caracterizados como objetos com uma dimensão física e uma localização espacial (Woodruff and Plaunt, 1994). Desta forma, a georreferenciação de texto pode ser associada à referência de um nome ou de uma morada num determinado texto, e a um local específico no planeta Terra que seja referido por essa expressão, podendo ser definido por um polígono ou por um centróide (centro de gravidade geoespacial), com uma latitude e uma longitude (Leidner, 2017).

Um sistema de georreferenciação de texto pode ser definido na área dos SIG como um sistema que utiliza técnicas de *geoparsing* (Leidner, 2017). Este termo é usado para descrever o processo de identificação automática dos locais referidos no texto (Gelernter and Balaji, 2013). No domínio da Informática, este sistema é conhecido como reconhecimento de entidades e classificação (Leidner, 2007). Segundo Gritta et al. (2018), *geoparsing* é a tarefa de identificar e resolver topónimos em relação às suas coordenadas geográficas. Segundo Infopédia (2021), um topónimo representa um nome genérico para um qualquer local ou entidade geográfica. Vulgarmente, junto com o *geoparsing* aparece a referência de um outro processo: o *geotagging*. Este processo é responsável por adicionar um conjunto de metadados contendo informação geográfica relativa a um determinado elemento, isto é, informação geográfica relativa aos nomes dos locais referidos no texto. O *geotagging* prepara a próxima fase da georreferenciação de um texto (Harvey, 2014). Além destes dois termos, no domínio dos sistemas de georreferenciação podemos encontrar um outro termo: *geocoding*. Basicamente *geocoding* é a tarefa de associar um conjunto de dados georreferenciados implicitamente a um conjunto de dados georreferenciados explicitamente, ou seja, representações com latitude e longitude, como acontece nas associações a mapas. Segundo McDonald et al. (2017) é o "processo de correspondência de códigos postais com as coordenadas geográficas correspondentes (latitude, longitude)".

2.2 UTILIDADE PRÁTICA

A georreferenciação de textos tem grande utilidade nos dias de hoje, uma vez que, com a ajuda da digitalização e da publicação online, esta facilita o acesso de utilizadores não especializados a informação geográfica, patrimonial, urbana ou ambiental de interesse. Através da georeferenciação podemos incorporar mapas atuais, como *Google maps*, *Bing maps*, bem como outros mapas, mais antigos, e fazer a sua comparação, promovendo o desenvolvimento de estudos históricos diacrónicos ou evolutivos por urbanistas, historiadores e arquitetos, sobre uma cidade, um estado ou um território, mostrando dessa forma a sua evolução ao longo do tempo (Cascón-Katchadourian et al., 2018). Além dos casos referidos, a capacidade de localizar geograficamente eventos em texto disponibiliza uma fonte valiosa de informação para todas as aplicações reais que se focam em respostas de emergências ou na análise de eventos geográficos nos média em tempo real (Gritta et al., 2018).

Historicamente, os documentos eram usualmente indexados por assunto, por autor, por título e, em menor grau, por tipo de documento. Com o passar dos anos, utilizadores de sistemas de informação começaram a requerer um acesso geograficamente orientado às coleções de documentos. Por exemplo, alguns administradores de recursos naturais que pretendiam recuperar informação pertinente em determinadas áreas, geólogos que procuravam localizar publicações que referiam locais, historiadores que pretendiam recuperar documentos sobre áreas específicas, ou mesmo jornalistas e turistas que pretendem localizar informação histórica sobre certas áreas (Woodru, 1994). Pretensões como estas deram origem à criação deste tipo de sistemas, tornando-os muito utilizados ao nível da georreferenciação de documentos históricos, uma vez que, permitem extrair informação espacial para analisar e visualizar usando sistemas SIG. Além disso, a combinação destes elementos facilita a descoberta de informação espacial que esteja contida em documentos. A descoberta deste tipo de informação permite, ainda, determinar as partes do texto que devemos ler de forma mais atenta (Rupp et al., 2013). No estudo de qualquer pesquisa histórica a compreensão da localização é uma parte crítica, bem como a compreensão de referências geográficas precisas e automáticas. A utilização deste tipo de informação permitiu aos historiadores descobrirem vários outros elementos de dados relacionados com regiões, para além do nome com que realizam a pesquisa. Se todas as coleções digitalizadas fossem georreferenciadas teríamos um unificador comum que nos permitiria agrupar diferenças e recursos (Grover et al., 2010).

Os sistemas de georreferenciação de textos permitem reduzir o tempo e o custo do processamento dos documentos e das coleções históricas. Isto é possível porque a informação espacial pode ser anotada manualmente, mas o tempo e custo necessários para o fazer tornam o processamento manual uma solução pouco viável para um conjunto elevado de dados (Leidner, 2017). Adicionalmente, a resolução de referências de locais individuais em documentos permite suportar a obtenção de informação geográfica Geographical Information Retrieval (GIR) que possa estar relacionada com algumas tarefas de processamento, como, por exem-

plo, a recuperação de documentos ou a visualização cartográfica de texto em documentos (Dias et al., 2012).

2.3 Utilização

Segundo Gregory et al. (2015), para se conseguir georreferenciar um texto é necessário identificar os diversos nomes de lugares que ele contém e, depois, atribuir as coordenadas geográficas que representam a sua localização. Este processo pode ser realizado em duas etapas. A primeira realiza-se utilizando técnicas de PLN que permitam reconhecer automaticamente os nomes de locais incluídos no texto. A segunda envolve o emparelhamento dos locais encontrados com os dados provenientes de um *gazeteer* (base de dados que incluí topónimos, tipos de recursos geográficos e pegadas espaciais).

Como método, a georreferenciação automática de texto é uma combinação de um processo NER, que reconhece e classifica nomes de locais no texto - *geoparsing* - com um processo de resolução de topónimos que desambigua os nomes de locais no texto usando contexto, quer linguístico, quer espacial, ou seja, seleciona a única interpretação espacial, provavelmente a mais ocorrente, entre um conjunto de locais candidatos num *gazetteer* (Leidner, 2017). Por fim, realiza-se a fase de *geocoding* (McDonald et al., 2017).

Os métodos de resolução de topónimos que descrevemos anteriormente podem ser classificados em duas categorias, segundo Leidner (2007):

- métodos baseados em heurísticas, que utilizam regras com base na intuição e observação humana;
- métodos baseados em métodos de aprendizagem, que têm por base associações estatísticas, ou seja, correlações adquiridas a partir do documento de treino e os recursos associados a partir do contexto em que ocorrem.

Quanto à fase da resolução de topónimos, esta inclui as seguintes etapas: identificação dos nomes dos locais referidos no texto, utilizando *geoparsing*, a fase de recuperação de informações de localização de locais candidatos, terminando na resolução de topónimos selecionando um local candidato no conjunto recuperado. A primeira fase, segundo Pouliquen et al. (2006), só ocorre depois de terem sido identificados nomes de pessoas e de organizações, que só serão identificados caso não tenham sido associados a outro tipo de nomes. Além disso, os nomes de locais são identificados exclusivamente através da consulta de *gazeteers*. Todas as fases referidas são executadas de forma sequencial (ou de forma integrada) na fase de resolução de topónimos (Leidner, 2017).

2.4 Ferramentas de georreferenciação

Nos últimos anos, no domínio da georreferenciação de texto, têm aparecido alguns trabalhos relevantes, além de existirem alguns produtos de *software* no mercado. Através de uma

pesquisa cuidada, identificamos seis sistemas (ferramentas) de georreferenciação que despertaram a nossa atenção. O *ArcGIS*, o *OpenCalais*, o *CLAVIN*, o *The Edinburgh Geoparser*, o *Geoparser.io* e o *GeoTXT*. De seguida, apresentaremos cada uma destas ferramentas de forma um pouco mais detalhada, permitindo perceber a importância e utilidade destes sistemas e ainda tirar proveito de algumas abordagens no desenvolvimento do sistema de georreferenciação para o *Tommi*.

2.4.1 *ArcGIS*

O *ArcGIS* (ESRI, 2020b) é um SIG criado especificamente com o intuito de trabalhar mapas e dados fornecidos e mantidos pelo Environmental Systems Research Institute (ESRI). Este *software* fornece um conjunto de funcionalidades bastante diverso aos seus utilizadores, nomeadamente guardar e gerir dados, criar mapas profissionais em 2D e 3D e disponibilizar meios de análise espaciais tradicional e avançada (ESRI, 2020b).

Além da sua versão *desktop* (Figura 1), o *ArcGIS* disponibiliza uma versão online, que permite ligar pessoas, localizações e dados utilizando mapas interativos. Tal como a versão *desktop*, a versão *online* permite a criação de mapas, a realização de análises de dados, a importação de dados do utilizador e a partilha dos mapas criados ESRI (2020c) com outros utilizadores.

Figura 1: Um dos ambientes de trabalho da Plataforma ArcGIS ESRI (2020a)

2.4.2 *OpenCalais*

O *OpenCalais* Refinitiv (2019) tem algumas semelhanças com a ferramenta anterior no que diz respeito ao processo de anotação. Embora a ferramenta anterior seja direcionada exclusivamente para a anotação de locais, o *OpenCalais* utiliza técnicas de PLN, de análise de textos e de mineração de dados, para derivar o significado de um elevado volume de conteúdo não estruturado de texto, anotando de forma rápida, fácil e precisa nomes de pessoas, lugares, factos e eventos (Refinitiv, 2019).

2.4.3 *CLAVIN*

O sistema *CLAVIN* (Novetta, 2019) é bastante idêntico aos anteriores. É um sistema de código fonte aberto que foi desenvolvido especialmente para a realização de tarefas de identificação e etiquetagem geográfica de documentos, executando estas tarefas baseados em contexto do texto. Tal como o *OpenCalais*, o *CLAVIN* não identifica somente os nomes dos locais mencionados num texto. Também utiliza heurísticas para determinar qual o local referido, com base no contexto do próprio documento. Além disso, o *CLAVIN* possui mecanismos de pesquisa difusa, que são capazes de manipular nomes de locais, com ortografia incorreta, e de reconhecer nomes alternativos como referência à mesma entidade geográfica. Este sistema permite ainda realizar pesquisas geoespaciais hierárquicas e análises geoespaciais avançadas em repositórios de dados não estruturados, uma vez que é capaz de enriquecer os documentos analisados com dados geográficos estruturados (Novetta, 2019).

2.4.4 *The Edinburgh Geoparser*

É um sistema que foi concebido para analisar um texto e identificar as ocorrências de locais, fixando essas ocorrências num mapa, e determinar a latitude e a longitude desses locais. Estas tarefas envolvem a realização de processos de desambiguação de nomes para os casos em que um dado local pode ter mais que uma interpretação possível, tal como acontece, por exemplo, com o país 'Peru' e a cidade indiana 'Peru', dois locais distintos com o mesmo nome. Além disto, este sistema é capaz de reconhecer automaticamente referências a lugares que estejam contidas em textos e desambiguá-los em relação a um *gazetteer*. O reconhecimento é necessário não só para identificar as referências a locais como também para a indicar a localização de possíveis entradas correspondentes no *gazetteer*, com valores de latitude e longitude, de modo que seja possível apresentá-los num mapa.

The Edinburgh Geoparser Alex et al. (2015) é um sistema que foi concebido para analisar o texto e identificar as ocorrências de locais, fixando-as num mapa, e determinar a latitude e longitude desses locais. Estas tarefas envolvem a realização de processos de desambiguação de nomes para os casos em que um dado local pode ter mais que uma interpretação possível. Além disto, este sistema é capaz de reconhecer automaticamente referências a lugares

que estejam contidas em textos e desambiguá-los em recorrendo a um *gazetteer*. O reconhecimento é necessário não só para identificar as referências de locais como também para indicar a localização de possíveis entradas correspondentes no *gazetteer*, com valores de latitude e longitude, de modo que seja possível apresentá-los num mapa (Alex et al., 2015).

2.4.5 *Geoparser.io*

O *Geoparser.io* é um sistema um pouco diferente daqueles que já apresentámos. Basicamente, este sistema é uma Representational State Transfer (REST) Application Programming Interface (API), que é capaz de identificar nomes de locais que estejam referidos num texto, desambiguando-os e devolvendo como resultado Geographic JSON (geojson). Ao contrário dos sistemas anteriores, que permitem realizar mapeamento dos locais encontrados, este sistema apenas realiza o *geoparsing* do texto, retornando as coordenadas geográficas dos locais que identifica, em formato de geojson, tal como podemos ver através da Figura 2. Apesar de já ter sido abandonado, este sistema é (ainda) bastante útil na criação de um sistema de georreferenciação.

```
{
  "type": "Feature",
  "geometry": {
    "type": "Point",
    "coordinates": [125.6, 10.1]
  },
  "properties": {
    "name": "Dinagat Islands"
  }
}
```

Figura 2: Exemplo de geojson

2.4.6 *GeoTXT*

O *GeoTXT* Karimzadeh et al. (2019) é um sistema que foi concebido com o objetivo de reconhecer e localizar geograficamente nomes de lugares em textos não estruturados. Para realizar estas tarefas, este sistema utiliza algoritmos de NER e um mecanismo de procura para indexar, classificar e recuperar topónimos. Com isso, consegue realizar um *geoparsing* bastante escalável. Adicionalmente, estes sistema também disponibiliza um conjunto de mecanismos de recuperação rápida e de personalização de topónimos, e uma API flexível e interoperável, que permite reconhecer nomes de lugares num texto e resolvê-los através de um *gazetteer* (Karimzadeh et al., 2019).

2.4.7 *Ferramentas de Processamento de Linguagem Natural*

O termo PLN é normalmente utilizado para descrever o processamento realizado pelos componentes de software ou hardware num sistema informático ao interpretar a língua falada ou escrita, isto é, a língua natural (Drouin, 2004). Este tipo de processamento oferece a vantagem de permitir que o computador consiga inferir e analisar a linguagem humana, ou seja, a linguagem natural, com mais significado do que apenas ler e responder a respostas programadas, oferecendo assim uma melhor capacidade de utilização e interação com os utilizadores.

Nesse leque de ferramentas de PLN podemos encontrar o *LinguaKit* (CILENIS SL, 2020). Esta é uma ferramenta de PLN, multilíngua, que integra vários módulos organizados em quatro secções distintas, nomeadamente: uma secção que inclui o módulo de conjugação de verbos ou o módulo de tradução para abordar os aspetos mais genéricos da linguagem; uma secção orientada para um perfil de utilizador do domínio educacional, que inclui os módulos de anotação morfossintática e o analisador sintático; uma terceira secção destinada a profissionais de comunicação e de marketing, que inclui o analisador de sentimentos e o extrator de palavras-chave; e por fim, uma quarta secção, experimental, na qual estão incluídos os módulos novos que foram acrescentados. Esta ferramenta pode ser utilizada via linha de comando ou através da plataforma online criada pelos seus construtores.

O LinguaKit disponibiliza um conjunto de serviços bastante interessante. De referir, um analisador de dependências (DepPattern), um marcador de Part-of-Speech tagging (PoS), o NER, Named Entity Classification (NEC), um serviço de resolução de correferência de entidades nomeadas, outro de análise de sentimentos, um extrator de palavras-chave e de relações, o reconhecimento de idiomas, a segmentação de frases, e *Lematização*. O *LinguaKit* realiza todas estas tarefas em português, espanhol, inglês, galego e português arcaico. Como tal, foi uma peça fundamental no desenvolvimento do nosso sistema de georreferenciação, dada a sua capacidade de realizar serviços NER, os quais representaram uma parte importante da anotação de nomes de locais inseridos nos textos que trabalhámos.

Outra ferramenta de PLN muito interessante é o *FreeLing* (Padró and Stanilovsky, 2012). Basicamente, uma biblioteca de programas, escritos em *C++*, que oferece uma interface em linha de comando, que pode ser usada para analisar textos. Os resultados dos processos de análise do *FreeLing* são comunicados em vários formatos, como, por exemplo, JavaScript Object Notation (JSON), Extensible Markup Language (XML) ou *ConLL*. O *FreeLing* foi criado com o objetivo de ser usada como uma biblioteca externa numa qualquer aplicação. Os seus serviços são bastante diversificados. Incluem, por exemplo, a partição de texto em *tokens*, a divisão de frases, a realização de análises morfológicas, o reconhecimento de palavras compostas, o reconhecimento flexível de múltiplas palavras, a divisão de contrações, a previsão probabilística de categorias de palavras desconhecidas, ou a codificação fonética, entre muitos outros. O *FreeLing* pode realizar todas essas tarefas em português, inglês, espanhol, francês, italiano, alemão, russo, norueguês, catalão, galego, croata, esloveno, asturiano e ga-

lês. O *FreeLing* possui uma vasta quantidade de idiomas em relação ao *LinguaKit*, tornando-o mais capaz na realização das tarefas que referimos (Padró and Stanilovsky, 2012).

2.5 Análise Comparativa das Ferramentas

	Geoparsing	Geocoding	Analise de Dados
ArcGIS	Não	Sim	Sim
OpenCalais	Sim	Não	Sim
CLAVIN	Sim	Não	Sim
The Edinburg Geoparser	Sim	Sim	Não
Geoparser.io	Sim	Não	Não
GeoTXT	Sim	Não	Não

Tabela 1: Comparação de características das ferramentas de georreferenciação estudadas

Na Tabela 1 podemos ver algumas das características base de um sistema de georreferenciação, que foram selecionadas para podermos comparar os diversos sistemas que estudámos anteriormente. Basicamente, esta tabela resume as funcionalidades base destes sistemas, indicando a sua capacidade (ou não) para realizar as tarefas de *geoparsing* e de *geocoding*. Além desses dois aspetos, incluímos também um terceiro elemento para comparação (análise de dados), que nos permite ver se o sistema pode ser utilizado (ou não) em atividades de análises de dados, como a análise geoespacial, que, como vimos, é muito importante na criação de qualquer sistema de georreferenciação.

De todos os sistemas apresentados aquele que mais se equipara ao tipo de sistema que foi desenvolvido no âmbito desta dissertação é o sistema *The Edinburg Geoparser*, uma vez que permite realizar os processos de *geoparsing* e de *geocoding*, que são essenciais para o sucesso do nosso sistema. Além disso, existem alguns trabalhos Grover et al. (2010) sobre a utilização desta ferramenta na análise de documentos históricos que são muito interessantes no ponto de vista do desenvolvimento dos trabalhos desta dissertação. Excluindo o sistema *ArcGIS*, todos os sistemas realizam a primeira fase do processo de georreferenciação de texto. Destaca-se aqui o sistema *Geoparser.io*; trata-se de uma REST API que realiza o processo de *geoparsing*, retornando geojson, diferenciando-se assim das abordagens dos sistemas anteriores.

Em termos de *geocoding* só duas das ferramentas estudadas é que o realizam, nomeadamente, *o The Edinburg Geoparser* e o *ArcGIS*. O *ArcGIS* é uma ferramenta bastante especializada na criação de mapas. Como tal, apresenta um melhor grafismo e uma maior qualidade na realização dessa tarefa. Embora essa qualidade não seja necessária em relação ao tipo de sistema que queremos desenvolver, este sistema tornou-se num bom caso para estudo no processo de melhoramento dos elementos de visualização de dados que incorporámos no sistema que desenvolvemos. Em termos de análise de dados, temos dois sistemas que incluem esse tipo de serviço. São eles o *CLAVIN* e o *ArcGIS*. Ambos os sistemas permitem realizar análises geoespaciais tradicionais e avançadas, embora um deles, *CLAVIN* , trabalhe com o

texto depois do processo de anotação e o outro, *ArcGIS* trabalhe com dados que constituem um mapa criado a partir da ferramenta em questão.

Quanto às ferramentas de PLN apresentadas, concluímos que o *LinguaKit* e o *FreeLing* são ferramentas muito idênticas, apresentando serviços bastante similares nesta área de trabalho. São duas ferramentas que foram desenvolvidas em ambientes académicos. A maior diferença entre elas verifica-se no número de idiomas com que trabalham. Ambas as ferramentas atuam ao nível de linha de comando. Aqui, a maior diferença verificou-se ao nível do PoS: o *LinguaKit* apresenta divisões de preposições no resultado final, mas o *FreeLing* não. Isso faz com que seja mais difícil manter a integridade do texto com o *LinguaKit*. Porém, esta desvantagem pode ser eliminada ao nível da sua implementação no sistema. Assim, qualquer uma destas ferramentas de PLN poderia ter sido utilizada neste trabalho de dissertação. Porém, optámos por utilizar o *LinguaKit*, devido ao grau de familiarização com a ferramenta já adquirido e à sua maior facilidade de implementação.

3

O SISTEMA DE GEORREFERENCIAÇÃO

3.1 GEOPARSING E GEOCONDING

Na área de SIG, um sistema de georreferenciação de texto pode ser definido, basicamente, como um sistema que utiliza técnicas de *geoparsing*. Como vimos, este termo é usado para descrever o processo de identificar automaticamente os locais referidos num determinado texto (Gelernter and Balaji, 2013). Este processo é feito recorrendo a um sistema de PLN, identificando assim as ocorrências de locais no texto em análise (Alex et al., 2015). No domínio da informática este é conhecido como reconhecimento e classificação de entidades (Leidner, 2007). Outro termo que também é usual é o *geocoding*, que consiste na tarefa de associar um conjunto de dados georreferenciados implicitamente a dados georreferenciados explicitamente, ou seja, representações de dados contendo valores de latitude e longitude tal como acontece nas associações a mapas. McDonald et al. (2017) especificaram um pouco mais a definição de *geocoding*, dizendo que é o "processo de correspondência de códigos postais com as coordenadas geográficas correspondentes (latitude, longitude)". Além destes processos, os sistemas de georreferenciação têm também que permitir a aplicação de métodos que sejam capazes de desambiguar nomes de locais em texto e que, consequentemente, possibilitem obter uma maior taxa de acerto na anotação de locais.

Este tipo de sistemas é muito útil na análise de documentos históricos, pois permite descobrir informações espaciais contidas nesses documentos, permitindo aos historiadores, aos geólogos, aos jornalistas e qualquer cidadão descobrir informação histórica sobre áreas de trabalho bem específicas (Woodru, 1994).

De forma que seja possível fornecer este tipo de informação, têm que ser criadas bases de dados georreferenciadas. A criação deste tipo de bases de dados é um grande desafio na criação de um sistema de georreferenciação, uma vez que necessitam de pesquisas demoradas acerca dos locais, de modo a serem adicionados corretamente, e um bom financiamento para a sua criação. Atualmente, tendo disponíveis essas bases de dados georreferenciadas, o trabalho no domínio dos sistemas de georreferenciação tornou-se mais fácil (Gregory et al., 2015).

Além da utilidade destes sistemas ao nível de documentos históricos, a sua capacidade de localizar geograficamente referências a locais em documentos textuais é uma grande mais valia, sendo muito importante para o fornecimento de informação em aplicações do mundo

real, como, por exemplo, em resposta a emergências, na análise de eventos geográficos nos meios de comunicação social, muitas vezes em tempo real, na compreensão de instruções de localização em sistemas de resposta automática, entre outros (Gritta et al., 2018).

3.2 Modelos e Processos de Georreferenciação

Atualmente, já podemos encontrar no mercado alguns tipos de sistemas de georreferenciação de texto. Alguns desses sistemas emergiram de projetos académicos. Por exemplo, o *Edinburgh Geoparser* Grover et al. (2010) foi utilizado em vários estudos para georreferenciar coleções históricas (Figura 3). Como podemos observar nessa figura, o sistema é composto por dois módulos. O primeiro é um *geotagger*. Este tem como função ler o ficheiro de texto e, através da utilização de PLN, obter a identificação dos locais referidos no documento.

Além deste primeiro módulo, podemos ver nessa figura um segundo, denominado por *georesolver*. Esse módulo tem como tarefa georreferenciar os locais encontrados, criando um ficheiro com os locais e as respetivas coordenadas geográficas. Tal como podemos observar na figura 3, este módulo recorre a um dicionário de locais (*gazetteer*) para obter as respetivas coordenadas geográficas (Grover et al., 2010).

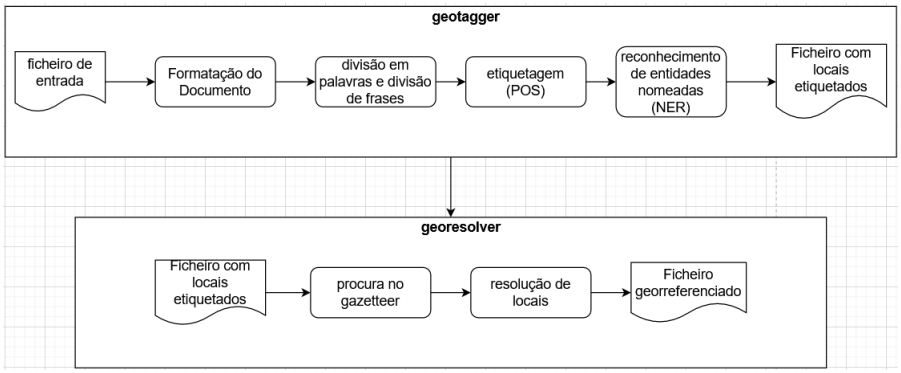


Figura 3: Modelo de Sistema de Georreferenciação adaptado de Grover et al. (2010)

Com o objetivo de entender o funcionamento de vários tipos de modelos, de modo a retirar o melhor de cada um e adaptar para o sistema de georreferenciação que queríamos criar, foi analisado em detalhe o trabalho de (Chen et al., 2018). Neste estudo, os autores tiveram como objetivo combater o problema da ambiguidade em processos de georreferenciação de texto. Através da utilização de grafos de locais, conseguiram resolver o problema da ambiguidade na anotação de locais. No diagrama de Business Process Modeling Notation (BPMN) representado na figura 4 podemos ver o processo de georreferenciação desenvolvido por (Chen et al., 2018). Este processo desenvolve-se em quatro fases. Na primeira fase faz-se a identificação de locais com base na procura num dicionário de locais, utilizando-se um algoritmo de aprendizagem não supervisionado, denominado por *Density-based clustering* (Webb et al., 2011), que identifica grupos (*clusters*) diferentes nos dados, considerando os pontos distintos como ruído (*outliers*) (Webb et al., 2011). Por último nesta fase existe um processo de desam-

biguação. Na segunda fase procuram identificar-se locais ou regiões aproximadas, isto é, locais aproximados dos que foram encontrados na fase anterior, de modo a integrar diferentes espaços de pesquisa e serem encontrados aproximadamente os locais que sobraram da fase anterior. De seguida, com os locais aproximados identificados, procuram-se correspondências no dicionário de locais e avaliam-se os locais encontrados, de modo a obter resultados o mais precisos possível.

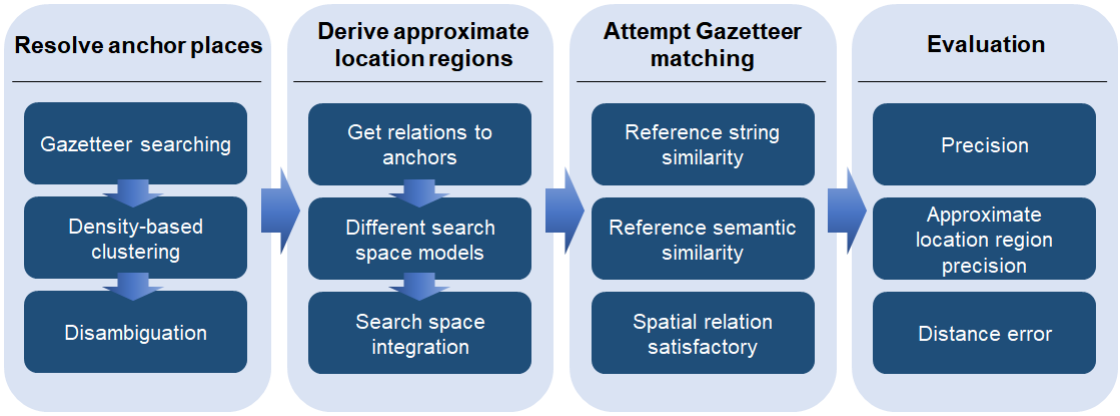


Figura 4: Abordagem do Modelo de Georreferenciação extraído de Chen et al. (2018)

Outro trabalho muito importante nesta área foi o de Leidner (2017), que foi desenvolvido especificamente para fazer a criação de um sistema de georreferenciação de texto. Neste estudo é apresentado um outro processo de georreferenciação de texto, no qual se identificam os locais referenciados num texto através da procura de referências num dicionário de locais. Neste trabalho apresentaram-se dois tipos de métodos para resolução de topónimos, ou seja, para a desambiguação de locais mencionados no texto, que utilizam, respetivamente, heurísticas ou processos de aprendizagem. Este trabalho foi importante para percebermos como podemos resolver questões relativas a nomes de locais que estejam expressos de forma ambígua num texto.

Além dos estudos referidos, foram estudadas outras ferramentas que estão disponíveis no mercado, como o Refinitiv (2019), que é uma ferramenta que utiliza técnicas de PLN para análise textos e mineração de dados, tendo como umas das funções anotar nomes de locais nos textos, além de anotar outras características presentes nos textos analisados. Outra ferramenta existente no mercado estudada foi o Novetta (2019), que tem como função realizar a identificação e etiquetagem geográfica de documentos. Por último temos o ESRI (2020b), que apresenta funcionalidades de *geocoding*, representando e editando coordenadas geográficas em mapas. Com este estudo realizado conseguimos adquirir uma grande variedade de conhecimento em relação a modelos de georreferenciação e em relação às fases que constituem o processo realizado pelos sistemas de georrreferenciação de texto.

3.3 O MODELO DESENVOLVIDO

Com o objetivo de criar um sistema de georreferenciação de texto para o Sistema *"Tommi"* (Barros et al., 2020), que permitisse identificar as referências a locais contidas nos textos do manuscrito *"O Livro das Propriedades do Tombo da Mitra"* Barros (2019) Barros (2021) e, depois, mapeá-los para observação num mapa, idealizámos e desenvolvemos um modelo idêntico àquele que foi demonstrado na figura 3 na secção 3.2. O modelo desenvolvido (Figura 5 incorpora, essencialmente, dois componentes distintos e interligados entre si, nomeadamente um *geoparser* e um *geocoder*. O *geoparser* vai ter como função encontrar todos os lugares existentes nos documentos e o *geocoder* vai mapear esses mesmos lugares, de modo que seja possível visualizar num mapa cada lugar referido nos documentos, oferecendo, assim, ao utilizador um maior conhecimento acerca dos locais encontrados.

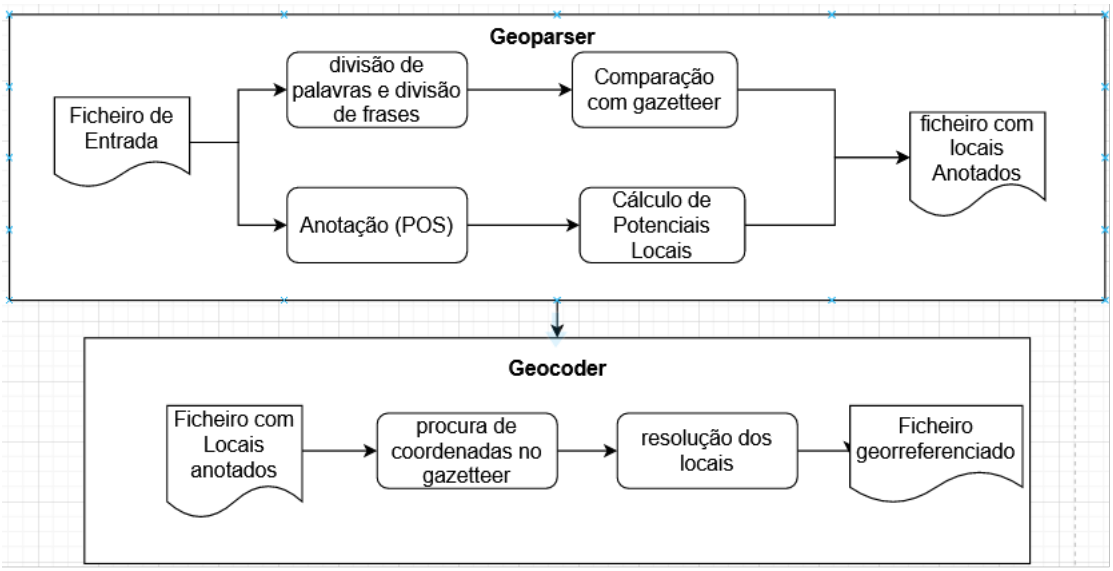


Figura 5: Esquema do Modelo do Sistema de Georreferencição

3.4 DESCOBERTA DE REFERÊNCIAS DE LOCAIS

O processo de descoberta das referências a locais nos textos é realizado pelo componente de *geoparsing* do sistema (Figura 6). Este componente é composto por dois módulos de identificação de locais, nomeadamente:

- O primeiro componente recorre a um *gazetteer*, que possui dicionários de nomes de locais, que incluem pegadas geoespaciais para os locais nomeados, e que são utilizados para identificar os nomes de locais do texto (Dunn, 2007)
- O Segundo componente recorre a identificadores de classe que possam estar presentes no texto, tais como ***Santa***, ***São***, ***Couto***, etc. A segunda diferença é efetuada no algoritmo NER, utilizado para identificar casos de paragem no texto.

Desta forma conseguimos aumentar o nível do reconhecimento dos locais identificados nos textos, além de tornarmos o processo de descoberta mais robusto 6.

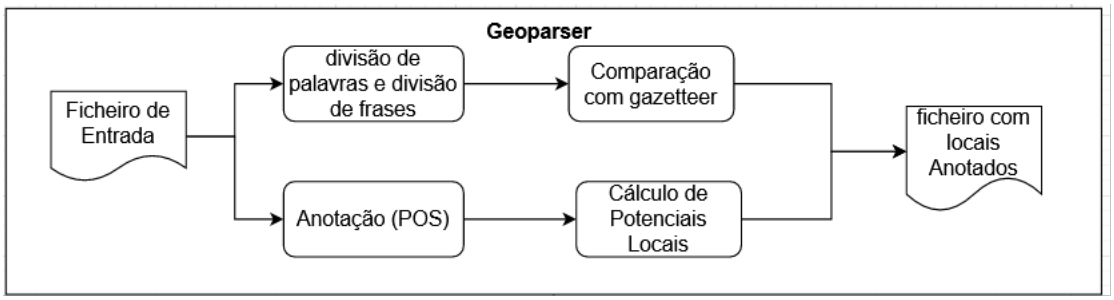


Figura 6: Esquema do Geoparser

Tal como foi explicado em cima, o primeiro módulo do *geoparser* criado identifica os nomes de locais no texto recorrendo a um dicionário de locais. Para isto é necessário realizar uma divisão de todo o texto em palavras e uma divisão por frases. Com esta divisão conseguimos realizar uma comparação de palavras com os locais do dicionário e ainda identificar nomes de locais com mais que uma palavra.

O segundo componente necessita primeiro que se realize a anotação e classificação de verbos, pronomes, determinantes no texto; com esta identificação realizada conseguimos identificar locais no texto com base nos identificadores de classe e nos respetivos casos de paragem de frases, como, por exemplo um sinal de pontuação ou um verbo. Depois da atuação deste módulo (Figura 6) ficamos com um ficheiro anotado da seguinte forma (Figura 7).

João Rodrigues de <*local* >***São Martinho*** </ *local* > e do poente com Diogo de <*local* >***Gouveia*** </ *local* > , e do norte com terras da Mesa de <*local* >***Esposende*** </ *local* > , e do sul com estrada de <*local* >***Leça da Palmeira*** </ *local* >;

Figura 7: Exemplo do Ficheiro com os locais Anotados depois de atuar *geoparser*.

Como podemos observar na figura 7, conseguimos identificar locais como ***São Martinho*** recorrendo ao segundo componente deste módulo e encontrar locais como: ***Esposende***, ***Leça da Palmeira***, ***Gouveia*** recorrendo à comparação das palavras no texto com o dicionário de locais.

3.5 Mapeamento de Referências

Após terminarmos a conceção do modelo de descoberta de referências de locais, passámos à construção do modelo necessário para fazer o mapeamento de referências, isto é, para suportar o processo de *geocoding* (Cascón-Katchadourian et al., 2018). Neste processo utilizámos o documento com os locais anotados anteriormente e obtivemos as suas coordenadas geográficas, utilizando o *gazetteer* referido anteriormente. Deste modo, obtivemos o documento georreferenciado, ou seja, uma associação das referências dos locais a posições concretas num

mapa, providenciado pelo *google maps*, *bing maps* ou outro qualquer software que seja capaz de lidar com mapas.

A fase de mapear as referências de locais pelo sistema de georreferenciação criado é evidenciada na figura 8.

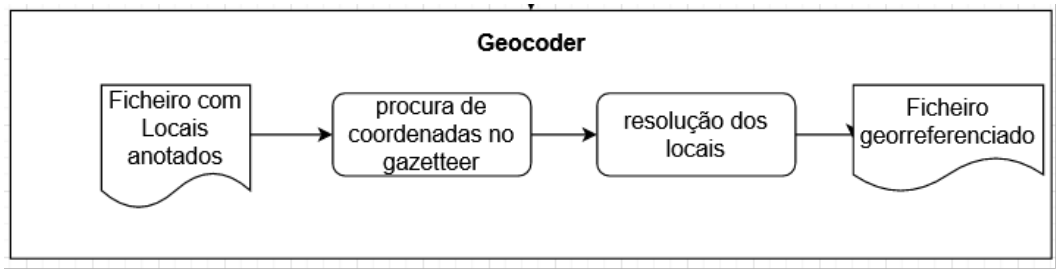


Figura 8: Esquema do *Geocoder*

Como podemos observar na figura 8, nesta fase o sistema de georreferenciação recorre ao documento anotado pela (Figura 6) e ao dicionário de locais que possuí além dos nomes de locais, as suas coordenadas geográficas. Com as coordenadas geográficas e com os locais encontrados, conseguimos mapear os mesmos.

4

CASO DE ESTUDO

4.1 APRESENTAÇÃO GERAL

O objetivo desta dissertação passa por implementar um modelo de georreferenciação e integrá-lo com o sistema *Tommi*. Esta dissertação surgiu da necessidade de oferecer aos estudiosos do *Livro de Propriedades* a possibilidade de encontrar os locais referidos nos documentos, bem como de mostrar esses mesmos locais num mapa. Estes duas necessidades, vem oferecer aos estudiosos uma melhor percepção acerca das regiões mencionadas nos documentos.

Para que os estudiosos utilizem este sistema vão precisar de aceder ao módulo de georreferenciação no sistema *Tommi* e processar os documentos já inseridos no sistema. Deste modo todos os locais são anotados e conseguem visualizar através de uma vista tabular ou uma vista num mapa, todos os locais encontrados, bem como um local individualmente.

Para a criação deste sistema de georreferenciação e do sistema *Tommi*, foi necessário o conhecimento e o devido tratamento dos documentos, provenientes do *Livro das Propriedades* do Tombo da Mitra.

Este sistema vai ser um módulo do sistema *Tommi* já implementado. Este módulo vai ter acesso aos documentos do sistema e vai realizar o seu processamento, armazenando todos os locais com as devidas coordenadas geográficas, de modo a conseguir marcar esses locais num mapa, permitindo a interação do utilizador com esse mapa, de modo a que consiga assim perceber em que cidade e distrito se insere essa localidade encontrada.

4.2 O LIVRO DAS PROPRIEDADES

De modo a definir o objetivo desta dissertação, têm de ser explicados dois pontos fundamentais, sendo eles o que é o *Livro de propriedades* e para que foi criado e consequentemente o que é o projeto *Tommi* e o objetivo do projeto.

O *Livro das Propriedades* ou *Tombo da Mitra* foi escrito no início do século XVII. Nesta altura as propriedades da Mesa Arcebispal encontravam-se por todo o Minho e Trás-os-montes, chegando mesmo ao bispado do Porto e a Santarém, bem como à Galiza (Barros, 2019) (Barros, 2021). Com o objetivo de se registarem todas essas propriedades, foi criado então o *Livro das Propriedades*. Este manuscrito tem registados, de uma forma extensa e pormenorizada,

as propriedades, as rendas e os foros da Mesa Arcebispal, tanto na zona da cidade de Braga como na região Norte de Portugal, estendendo-se para fora da mesma, tal como explicado anteriormente. Neste livro todos os fólios – um fólio corresponde a uma folha, ou seja, duas páginas – estão rubricados, tendo o códice sido devidamente encerrado e assinado em 1606. Ao longo dos 644 fólios que fazem parte do *Livro das Propriedades* estão apresentadas imensas referências a tipos de terras (agrícolas, de mato, de pasto, etc.), acidentes de terreno e muitas outras referências a aspetos geográficos (penedo, rego, fonte, ermida, quebrada, outeiro, outeirinho, etc) que na altura eram referidos. Além disto, podem encontrar-se, também, nomes de ruas, lugares, rios, povoações, proprietários, apontamentos biográficos e genealógicos, entre muitas outras coisas (Barros, 2019) (Barros, 2021).

Para que se possa fazer o estudo geográfico, sociocultural, agrícola, económico, arquitetónico e religioso de Braga, do Minho e de outros territórios portugueses alcançáveis pela Mesa Arcebispal de Braga, é muito importante a edição do manuscrito. A sua transcrição tornará possível a criação de uma ampla fonte de informação, muito rica, incluindo nomes de terras e respetivos emprazadores ou cultivadores. Será ainda importante para obter genealogias das famílias nas terras que o livro alcança.

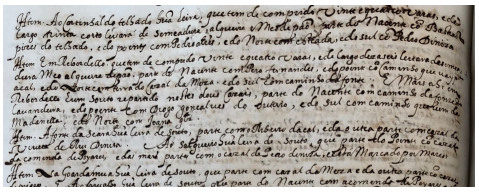
O *Livro de Propriedades* apresenta todas as propriedades como já referido anteriormente, que pertenciam a mesa Arcebispal de Braga, detalhando com grande promenor e precisão propriedades rústicas e urbanas das comarcas de Valença, Vila Real, Chaves e Braga e abrangendo terras pertencentes à Galiza. Nestes documentos temos acesso, além das informações já referidas, a dados das rendas ou pagamentos efetuados ou em dívida dos emprazadores à mesa Arcebispal (Barros, 2019) (Barros, 2021).

Para se ter uma ideia um pouco mais concreta do *Livro de Propriedades* e da sua dimensão, na Figura 9, apresentamos uma imagem do "Livro de Propriedades"fechado e do *Livro de Propriedades* aberto a meio. Através dessas imagens podemos ver a dificuldade que oferece a pesquisa no *Livro de Propriedades* e a dificuldade na edição do códice, o que vem mostrar a importância de um sistema que consiga processar o texto contido nos documentos do *Tombo da Mitra*, pesquisar sobre os mesmos e extrair informações relevantes para os estudiosos deste tema.

Figura 9: *Livro de Propriedades fechado e Livro de Propriedades aberto*

Tal como se pode observar pelas fotografias acima apresentadas cedidadas por Anabela Barros (Figura 9), podemos concluir que o *Livro das Propriedades* trata-se de um grande e pesado códice, que contém tal como referido anteriormente, uma extensa e pormenorizada relação das propriedades e rendas ou foros da mesa Arcebispal (Barros, 2019) (Barros, 2021).

Na figura 10 apresenta-se um pequeno excerto de um fólio pertencente ao *Livro das Propriedades*, escrito em português clássico, e a edição semidiplomática do mesmo, de Anabela Barros. Através deste exemplo podemos constatar que não é fácil ao utilizador comum aceder ao conteúdo de um fólio, mesmo quando se encontra em bom estado. A leitura e edição de manuscritos são processos lentos e complexos, uma vez que lidamos com a língua de outros períodos, hábitos de escrita diversos, ampla variação ortográfica e estados muito variados do suporte físico e da tinta.



Ittem. Ao Cortinhal do telhado hua' Leira, que tem de comprido vinte e quatro varas, e de / largo trinta e oito leuará de semeadura, alqueire e Meo de pão, parte do Nacente co' Bastião / pires do telhado, e do poente com Pedro alres', e do Norte com estrada, e do sul co' Pedro Diniza

Ittem Em Rebordello, que tem de comprido vinte e quatro varas, e de largo dezaseis leuara de semeadura / Meo alqueire de pão, parte do Nacente com Pero fernandes, e do poente co' Caminho que uay pª a / cal, e do Norte com terra do Cazal da Meza. e do sul com caminho da fonte: E mais ahi em / Rebordello hum souto repartido nestes dous Cazais, parte do Nacente com caminho da fonte da / lauandeira, e do poente com Diogo goncalues

Figura 10: Excerto do Folio 97 e sua transcrição

O estudo do *Livro das Propriedades* apresenta elevado interesse devido à informação que contêm, tal como as propriedades que pertenciam à mesa Arcebispal de Braga, os tipos de terrenos, o que se cultivava nesses terrenos ou o valor dos mesmos, sendo possível através do estudo dos documentos criar árvores genealógicas com os dados dos donos dos terrenos. Devido importância do conteúdo do Livro surgiu o interesse de disponibilizar os documentos num sistema capaz de catalogar, indexar, armazenar e explorar estes documentos, uma vez que o livro é de difícil acesso, visto ser único e só poder ser consultado no local em que está armazenado, impossibilitando o estudo a mais que uma pessoa ao mesmo tempo e a estudiosos que não se consigam deslocar ao Arquivo Distrital de Braga para o estudar.

No âmbito do projeto de disponibilização do texto editado do manuscrito no sistema, surgiu o interesse de georreferenciar o seu conteúdo, permitindo assim aos estudiosos explorar o *Livro das Propriedades* de uma forma interativa e didática, uma vez que oferece muita informação sobre a região norte de Portugal, bem como algumas povoações até Santarém e outras da Galiza, em Espanha.

Nas secções seguintes será apresentado o sistema *Tommi* e descrever-se-ão os serviços da aplicação, bem como a integração dos serviços de georreferenciação.

4.3 TOMMI

O sistema *Tommi* (tommi.di.uminho.pt) Barros et al. (2020) está a ser desenvolvido desde há cerca de um ano e meio no Departamento de Informática da UM. É um projeto de desenvolvimento interno coordenado pelo Professor Orlando Belo, do Departamento de Informática, e pela Professora Anabela Barros (autora da ideia base do projeto), do Departamento de Estudos Portugueses e Lusófonos, que tem estado a ser trabalhado em várias componentes de projeto de unidades curriculares de ensino e dissertações de mestrado. A idealização e construção do sistema teve como base a ideia de disponibilizar uma ferramenta computacional que permitisse fornecer a qualquer utilizador e aos diversos estudiosos do *Livro das Propriedades* Barros (2019) Barros (2021) meios para acederem ao seu conteúdo, de uma forma simples, fornecendo partes dos seus textos, sem que para isso se tenha de recorrer, repetidamente, à leitura dos documentos originais ou à consulta da edição semidiplomática em português seiscentista, ou mesmo da edição interpretativa (com ortografia atualizada).

O sistema *Tommi* vem assim oferecer uma forma moderna, simples e eficaz de estudar os textos contidos no *Livro das Propriedades*. O sistema está acessível online e pode ser acedido através do endereço 'tommi2.di.uminho.pt/admin'. Como tal, é um sistema que pode ser utilizado a parti de qualquer lugar, a qualquer momento.

O trabalho desta dissertação vem complementar as funcionalidades gerais do sistema, uma vez que irá facilitar os processos de análise de textos contidos no mesmo, incorporando um sistema de anotação automática, direcionado especificamente para a pesquisa de locais referidos ao longo do "Livro das Propriedades". O sistema de georreferenciação de texto providenciará um conjunto de meios que ajudarão a encontrar os locais que são referidos nos seus textos de forma direta ou indireta, permitindo com isso realizar a sua anotação de forma automática e posicionar num mapa, quando necessário, os vários locais encontrados nos textos, o que permitirá aos estudiosos ter uma ideia concreta da localização desses locais.

4.3.1 *Descrição do Sistema*

O sistema "Tommi" é um produto de software multi-plataforma, que foi implementado de forma a poder funcionar nos sistemas operativos mais relevantes. É um sistema com uma arquitetura típica cliente-servidor, na qual os serviços de processamento de dados e trata-

mento das estruturas de gestão de dados estão integrados na parte do servidor (backend) e os serviços relacionados com a interação com o utilizador e com o carregamento e pesquisa de documentos na parte cliente (frontend).

Figura 11: Ambiente de Autenticação do Sistema

Através do endereço indicado, podemos aceder à página de autenticação do sistema (Figura 11). Para além das funcionalidades mais comuns neste tipo de página (informação sobre o sistema, o seu desenvolvimento, questões de privacidade, etc.), nesta página podemos apresentar as nossas credenciais de acesso (utilizador, palavra-chave), previamente atribuídas pelo administrador do sistema, e aceder aos serviços que o sistema nos pode fornecer. Com a apresentação de um conjunto de credenciais válidas, o sistema transporta-nos para o seu ambiente de trabalho principal (Figura 12).

Figura 12: Ambiente Principal de Trabalho no Sistema

A partir deste ambiente de trabalho temos acesso aos serviços do sistema. Desse leque de serviços, salientamos os seguintes:

- Importação de Textos
- Gestão de Documentos
- Gestão de Índices
- Gestão de Utilizadores

Para realizar a inserção de um novo documento na base de dados (considerando separadamente cada meio fólio como um documento, ou seja, 1 rosto, 1 verso, etc., de forma a fazer corresponder cada fotografia do códice a cada parte do texto, ainda que, em muitos casos, incompleta), o utilizador pode aceder no menu à opção de **Importação**. A importação de documentos no sistema é realizada em seis passos, sendo o primeiro passo a catalogação do documento a inserir. Neste passo o utilizador submete o ficheiro, bem como os dados necessários para a identificação do documento, tal como o nome do mesmo (ou seja, o número do fólio e o lado, rosto ou verso, já que o manuscrito se acha foliado, e não paginado), e opcionalmente pode submeter uma foto do documento e acrescentar algumas observações (Figura 13).

Figura 13: Passo 1 da Importação - Catalogação

Acabada a catalogação do documento, o passo seguinte do processo de importação é a pré-visualização do documento inserido na Figura 14

Figura 14: Passo 2 da Importação - Visualização

O passo três é igualmente uma pré-visualização mas desta vez da fotografia inserida, caso tenha sido introduzida alguma.

O utilizador, ao confirmar que as pré-visualizações vão ao encontro do que ele pretende inserir chega ao quarto passo da importação de documentos, no qual o sistema identifica as etiquetas contidas no documento inserido, mostrando o resultado deste processo (Figura 15).

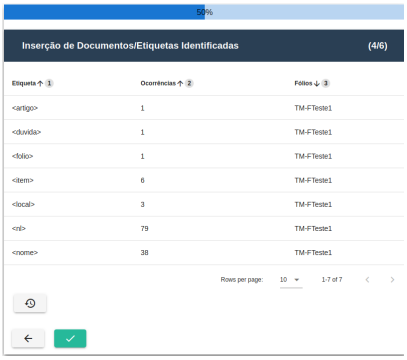


Figura 15: Passo 4 da Importação - Etiquetas Identificadas

O penúltimo passo da importação de documentos no sistema é a visualização do resultado do cálculo de índices. Este cálculo caso o utilizador termine o processo de importação, também é armazenado na base de dados do sistema, podendo ser posteriormente consultado (Figura 16).

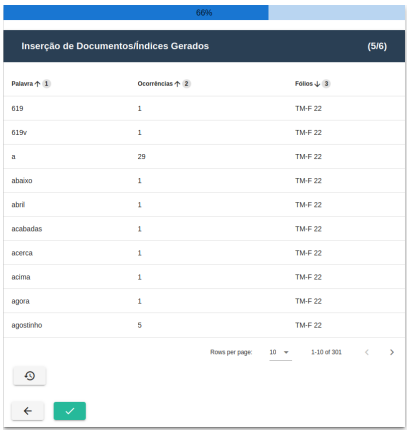


Figura 16: Passo 5 da Importação - Visualização de Índices

No último passo deste processo de importação é apresentado um resumo do que foi inserido.

O utilizador, através do menu de navegação, pode selecionar a opção de **Gestão de Fólios**. Nesta página o utilizador pode consultar todos os documentos armazenados no sistema, podendo ainda visualizar documento a documento, eliminar do sistema fólios inseridos e ver a fotografia caso tenha sido importada a imagem correspondente do manuscrito (Figura 17).

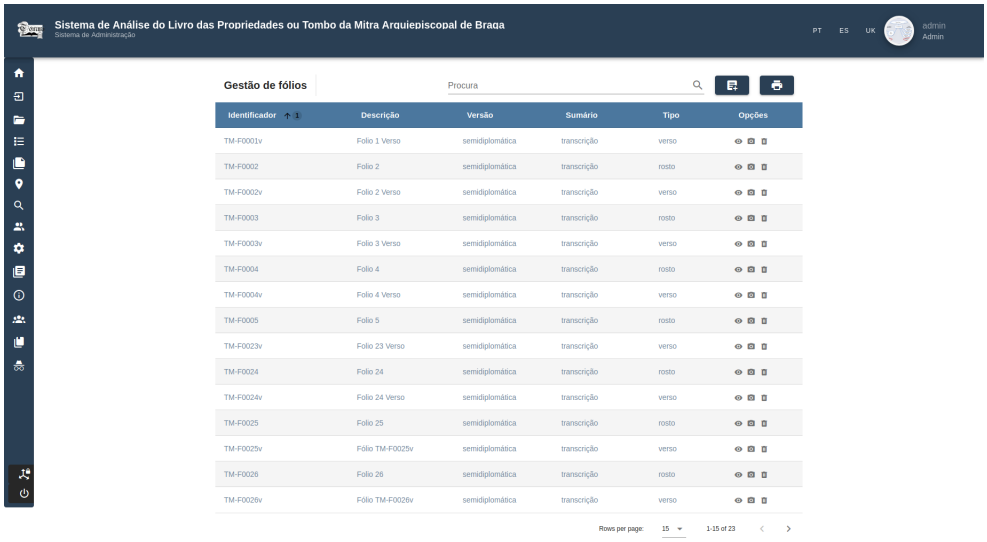


Figura 17: Gestão de Fólios

Além de poder realizar a gestão de fólios, o utilizador pode ainda gerir os índices gerados no momento de inserção do documento. Nesta página é possível consultar todos os índices do sistema e, tal como na gestão de fólios, o utilizador pode visualizar um índice individualmente (Figura 18).

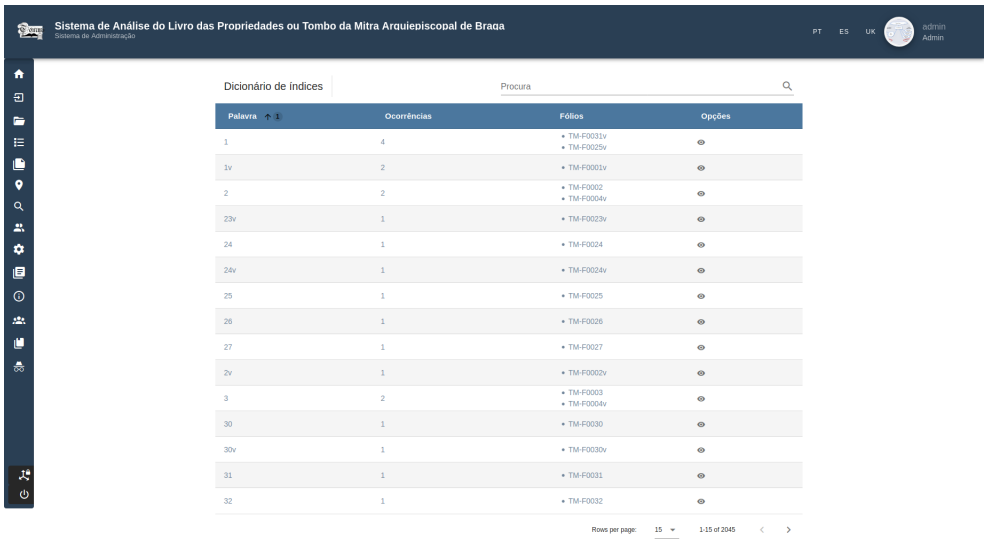


Figura 18: Gestão de Índices

Outra opção a que se pode aceder através do menu de navegação é a gestão de utilizadores. Esta opção tal como as anteriores, é uma das mais importantes no sistema, oferecendo ao utilizador visualização das informações de todos os utilizadores do sistema, podendo ainda eliminar um utilizador ou editar a informação do mesmo (Figura 19).

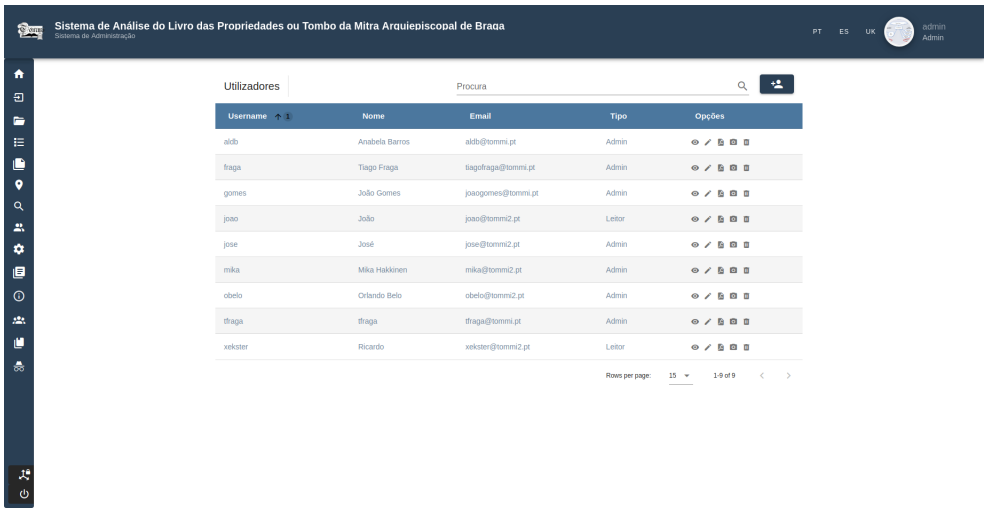


Figura 19: Gestão de Utilizadores

4.3.2 *A Integração de Serviços de Georreferenciação*

Durante os últimos dois anos, o sistema evoluiu significativamente, sofrendo mesmo grandes alterações na sua estrutura base e nas suas estruturas de interface com o utilizador (Barros et al., 2020). Além disso, em particular no último ano de desenvolvimento, emergiram novas ideias de serviços especialmente orientadas para o enriquecimento dos textos acolhidos pelo

sistema, envolvendo novos elementos de dados e mecanismos para análise dos seus conteúdos que implicaram a criação de novas áreas de trabalho no sistema. Como consequência apareceram as áreas de anotação e de georreferenciação de textos. Nesta dissertação, como já referimos anteriormente, abordaremos os últimos serviços referidos.

Os serviços de georreferenciação introduzidos no sistema *Tommi* consistem no processamento dos documentos inseridos no sistema, identificando e armazenando todos os locais encontrados, oferecendo ao utilizador capacidade de visualizar todos os locais. Além deste processamento e armazenamento, estes serviços vão permitir ao utilizador visualizar os locais em mapas devidamente identificados no mesmo, e ainda oferecer a visualização individual dos locais. Como as outras opções de gestão no sistema, estes serviços vão permitir ainda a remoção dos locais, caso o utilizador ache pertinente.

Com a inclusão destes serviços no sistema *Tommi*, pretendeu-se oferecer aos estudiosos do *Livro de Propriedades* capacidade de perceber de uma forma mais dinâmica e visual os locais referidos nos documentos, permitindo aos utilizadores conhecer a região que engloba os locais referidos, bem como as localidades próximas. Estes serviços, vêm melhorar a experiência de estudo destes documentos, bem como a experiência de uso do sistema criado, mostrando algo diferente além de listas e documentos textuais.

O processo de integração e todas as funcionalidades destes serviços de georreferenciação vão ser explicados em detalhe no capítulo 5, onde será descrito em pormenor todo o processo de implementação dos serviços de georreferenciação e ainda a integração com este sistema.

5

IMPLEMENTAÇÃO

O sistema de georreferenciação de texto implementado é constituído por dois elementos principais, o módulo de *geoparsing*, que tem como função identificar os nomes de locais no texto, e o módulo de *geocoding*, que é responsável por associar um conjunto de dados georreferenciados implicitamente a dados georreferenciados explicitamente, ou seja, representações com latitude e longitude, como acontece na associação a mapas. Na Figura 20, podemos observar de forma geral o modo como o processo de georreferenciação de texto é realizado. De seguida, os módulos referidos serão apresentados e descritos de forma mais elaborada.

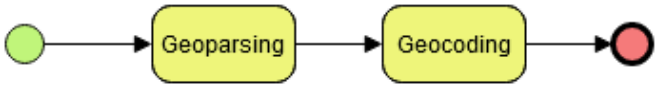


Figura 20: Diagrama do Processo realizado pelo sistema de georreferenciação de texto

5.1 CRIAÇÃO/PROCESSAMENTO DO DICIONÁRIO DE LOCAIS

Para que fosse possível georreferenciar os locais que estavam referidos nos textos editados a partir dos fólios do *Livro de Propriedades*, tivemos que criar um *gazetteer*, isto é, um dicionário de locais. Este dicionário foi construído com base no *Geonames*, que é uma base de dados geográfica que abrange todos os países do mundo. De modo a facilitar a extração da informação pretendida utilizámos um *dataset* (Tabela 2) em formato Comma Separated Values Files (CSV), com cerca de **1.37GB**.

geonameid	name	latitude	longitude	country code	modification date
2261580	Vila Verde	39.73545	-8.36266	PT	2011-08-28
2261581	Vila Verde	39.32603	-9.0878	PT	2011-08-28
2261582	Vila Verde	39.15489	-9.11512	PT	2012-01-17
2261583	Vila Verde	38.83333	-9.36667	PT	2011-08-28
2261584	Vila Velha de Ródão	39.65646	-7.6767	PT	2011-08-28
2261585	Vila Seca	39.12191	-9.15827	PT	2011-08-28
2261586	Vilas de Pedro	39.96891	-8.24624	PT	2011-08-28
2261587	Vilãs	38.92972	-9.3316	PT	2011-08-28
2261588	Vila Ruiva	38.24664	-7.93753	PT	2011-08-28

Tabela 2: Exemplo de alguns campos presentes nos dados do *Geonames*

O *dataset* que foi escolhido tem uma dimensão considerável, o que levantou alguns problemas no seu processamento, dados os recursos tecnológicos disponíveis. Para que fosse possível fazê-lo de forma eficiente, foi realizada uma fragmentação do ficheiro, de forma a obtermos apenas a parte relativa às localidades de Portugal. Depois desta fragmentação ficámos com um ficheiro CSV com todas as localidades de Portugal, permitindo assim iniciar o processamento do mesmo.

O processamento do ficheiro foi realizado através da criação de um *script* escrito em *Python*. O *script* começa por ler o ficheiro de localidades obtido. A partir da leitura desse ficheiro faz-se a criação de uma estrutura de dados específica, com capacidade para armazenar todas as referências de locais que pertencem a Portugal, e a filtragem das localidades que não têm interesse para o processo de georreferenciação que queremos realizar. O processo responsável pela criação do dicionário de locais está representado na Figura 21

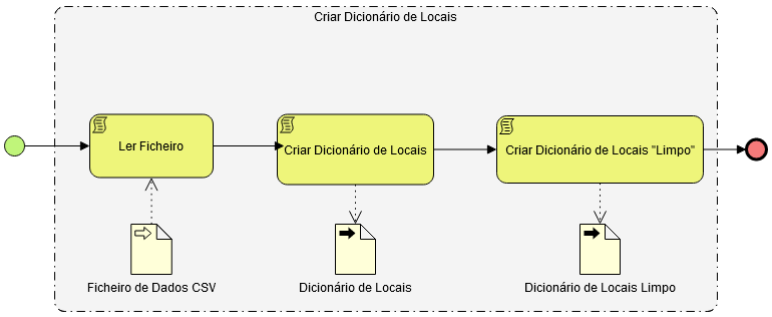


Figura 21: O Processo de Criação do Dicionário de Locais

Como podemos observar na Figura 21, o dicionário de locais é bastante importante na realização das fases seguintes do processo, como o *geoparsing* e a fase de *geocoding*, uma vez que o dicionário de locais possu a informação detalhada das localidades com interesse, o que inclui as suas coordenadas geográficas. Este processo inicia-se com a leitura de um ficheiro de dados em formato CSV (Ler Ficheiro). Os dados extraídos desse ficheiro são de seguida

transformados e colocados num ficheiro de dados JSON (Criar Dicionário de Locais). No processo de criação do dicionário de locais é criado ainda um outro ficheiro também em formato JSON, contendo os nomes dos locais em letra minúscula e com todos os acentos removidos (Criar Dicionário de Locais "Limpo"). Esta última tarefa é importante para a fase de *geoparsing*, uma vez que faz a homogeneização dos nomes dos locais, aumentando assim a precisão de acerto na sua anotação. Isso é necessário já que podem surgir locais com acentos em falta ou com letra minúscula em vez de maiúscula. Tendo estas tarefas terminadas, ficamos com dois ficheiros com todas as localidades de Portugal que interessam para o sistema de georreferenciação. O primeiro destes ficheiros conserva o nome correto das localidades de Portugal, enquanto o segundo mantém as localidades com o nome modificado, de modo a ser possível abranger as localidades referidas nos documentos, no caso de existirem variantes ortográficas dos topónimos.

5.2 ATUALIZAÇÃO DE GRAFIA

O módulo de atualização da grafia é um dos mais importante (e críticos) do sistema que implementámos. Tal deve-se ao facto de os documentos que fazem parte do Livro das Propriedades terem sido redigidos em português clássico e de o dicionário de locais que foi criado ter os seus dados expressos em português moderno. Isto faz com que a comparação de locais necessária à georreferenciação das localidades contidas no Livro das Propriedades não seja de fácil execução. Por isso, precisámos de criar no sistema um processo especificamente orientado para a atualização gráfica dos documentos (Figura 22) que queríamos processar, tanto quanto possível, de forma automática.

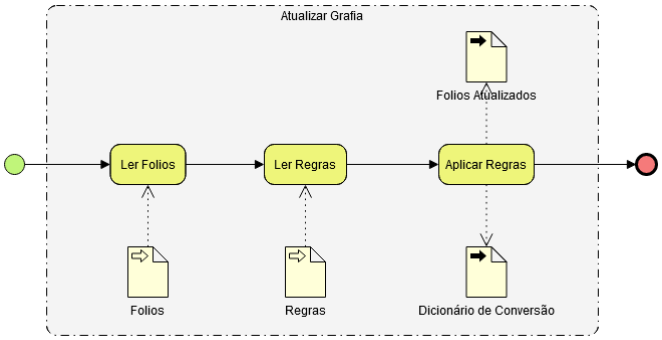


Figura 22: O processo de atualização de grafia

Como se pode observar na Figura 22, o processo de atualização de grafia inicia-se com a leitura dos textos do Livro das Propriedades que estão contidos no sistema (Ler Folios). De forma a preparar o processo de atualização de grafia, para além da tarefa de leitura dos fólios, foi necessário criar um conjunto de regras de atualização (Figura 23). Tal como se pode observar, as regras de atualização que foram criadas apresentam a palavra ou a expressão antiga (tal como aparece nos documentos) e a sua (devida) substituição.

Figura 23: Exemplos de regras de atualização de grafia

As regras que foram criadas são lidas durante o processo de atualização (Ler Regras) para que possam ser aplicadas sobre o conjunto de fólios do sistema (Aplicar Regras). Esta última tarefa do processo de atualização de grafia tem como resultado um conjunto de fólios devidamente atualizado e um dicionário de conversão. O dicionário de conversão (Tabela 3) possuirá todas as palavras dos documentos processados, por exemplo "Sendo/Sendo". Neste caso, isto quer dizer que a palavra "Sendo" não foi afectada por nenhuma regra de conversão. Mesmo nestes casos as palavras são armazenadas no dicionário de conversão. Noutro exemplo, a palavra **"dyto"** deu origem à palavra **"dito"**, tendo isso sido registada no dicionário de conversão e nos documentos modernizados.

Sendo	passado	carta	para	o	dyto	Afonso
Sendo	passado	carta	para	o	dito	Afonso

Tabela 3: Dicionário de Conversão

Finalizada a atualização de grafia e a criação do dicionário de locais, o sistema de georreferenciação pôde passar à realização dos processos de *geoparsing* e *geocoding*.

5.3 ANOTAÇÃO DE LOCAIS

O primeiro processo realizado pelo sistema de georreferenciação de texto é o *geoparsing*. Como já referido, este processo tem como objetivo fazer a anotação dos nomes dos locais que estão referidos nos textos. Para que fosse possível encontrar esses locais, necessitámos de implementar os dois módulos anteriores: criação do dicionário de locais e atualização de grafia. A Figura 24 mostra o processo de *geoparsing* em detalhe, bem como as diversas tarefas que nele são executadas.

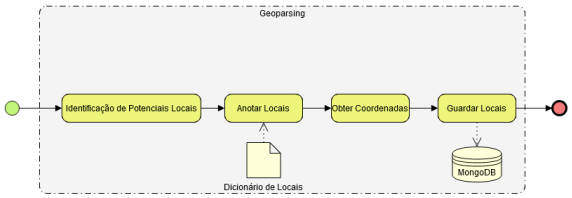


Figura 24: O processo de *geoparsing* do sistema de georreferenciação de texto

O processo de anotação de locais (*geoparsing*) considera a realização de vários subprocessos, nomeadamente a identificação de elementos de texto que possam ser potenciais locais (Identificação de Potenciais Locais), a anotação de locais com base no dicionário (Anotar Locais), a obtenção de coordenadas geográficas do dicionário de locais (Obter Coordenadas) e o armazenamento na base de dados dos documentos que contêm a localidade encontrada e as suas coordenadas (Guardar Locais).

5.3.1 *Identificação de Potenciais Locais*

A construção do módulo de atualização de grafia veio facilitar o desenvolvimento do (sub)processo de identificação de (potenciais) locais. O *Livro das Propriedades* é um manuscrito de inícios do século XVII. Como tal, o português utilizado também é seiscentista (clássico), contendo palavras diferentes ou registadas numa ortografia variável e em muitos casos distinta da atual. Além disso, nos textos do livro existem localidades que já não se conhecem por esses nomes ou não fazem parte de Portugal. Por esses motivos, tivemos que fazer a atualização da grafia e criar um módulo específico para fazer a identificação dos locais que não se conseguem anotar com base no dicionário que temos disponível.

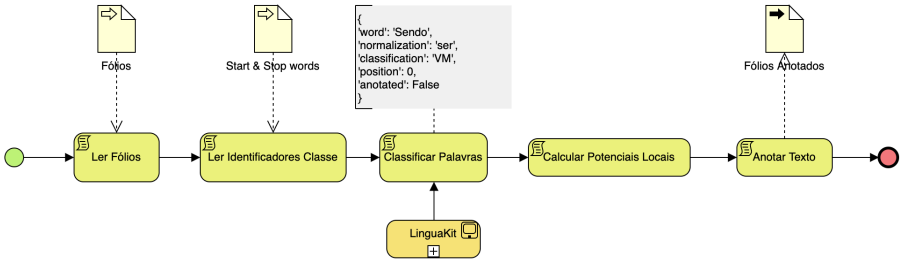


Figura 25: O processo de identificação de Potenciais Locais

O processo de identificação de potencias locais está apresentado na Figura 25. Tal como podemos ver no diagrama apresentado, a estratégia de identificação de potenciais locais começa por fazer a leitura dos textos dos fólios. De seguida, procede-se à obtenção dos identificadores de classe que tenham sido identificados, à classificação das palavras em questão e à determinação dos potenciais locais. Por fim realiza-se a anotação do texto, guardando-se de seguida os textos devidamente anotados.

Identificador de Classe
Couto
Santa
São
Aldea
Vila

Tabela 4: Identificadores de Classe

Para fazermos a criação deste módulo começámos por determinar os vários identificadores de classe – vejam-se alguns exemplo na tabela 4. Como podemos verificar, os identificadores de classe são palavras que estão contidas no texto que, por norma, antecedem ou pertencem a localidades. Estes identificadores de classe permitem revelar que, possivelmente, juntamente com eles costuma existir um nome de uma localidade. O passo seguinte da estratégia é dado no sentido de identificar casos de paragem ao longo do texto, isto é, palavras que sejam determinantes, adjetivos ou verbos, que possam marcar o fim do nome ou da pesquisa, caso não seja encontrada nenhuma referência a uma localidade. Na tabela 5, podemos observar os critérios de paragem que foram identificados.

Critério de Paragem	
Simples	**Especial**
Verbos	Preposições
Nomes Comuns	
Determinantes	
Adjetivos	
Advérbios	
Conjunções	
Pontuação	

Tabela 5: Critérios de Paragem

Depois de terminado o processo de identificação dos critérios de paragem e dos identificadores de classe, o processo de identificação de potenciais locais começa por fazer a leitura dos ficheiros JSON, relativos aos critérios de paragem e aos identificadores de classe, e armazena-os em estruturas de dados adequadas. Como se pode verificar pelo diagrama da Figura 25, de seguida, é executada uma função de classificação que tem como objetivo organizar as palavras dos textos por categoria, nomeadamente, em nomes, adjetivos, verbos, advérbios, etc. Para realizar esta classificação foi utilizada a ferramenta *LinguaKit* (CILENIS SL, 2020). É esta ferramenta que realiza a tarefa de reconhecimento de entidades (NER). A partir do texto classificado, o processo consegue encontrar os casos de paragem presentes nos documentos e, assim, determinar potenciais locais para anotar. Como resultado, obtemos um texto anotado, no qual podemos encontrar anotações como a seguinte, relativa a um potencial local:

- Mesa Arcebispal por dia de <*local* >***São Miguel*** <*/ local* > de cada ano ao Sr. Arcebispo;

Como podemos ver neste pequeno exemplo, o local anotado nesta frase, **São Miguel**, foi identificado com base nos identificadores de classe, ou seja, nas palavras que marcam o início do calculo, e nos critérios de paragem definidos sejam, embora neste caso pelo contexto São Miguel refere-se ao dia do santo e não ao local.

5.3.2 *Anotação de Locais com base no dicionário de locais*

O módulo de anotação de locais (Figura 26) é um pouco menos complexo que o módulo de identificação de potenciais locais (Figura 25), uma vez que integra um conjunto de tarefas um pouco mais simples. Tal como se pode observar pelo seu diagrama de processo, o módulo de anotação de locais inicia a sua atividade com a leitura dos textos dos fólios, recorrendo ao dicionário de locais criado no processo anterior, e depois analisa todas as referências de locais, comparando-as com as palavras contidas no texto. De forma a evitar erros na escrita dos locais nos documentos, o processo trabalha os textos sempre em minúsculas e sem quaisquer acentos, usando o dicionário de locais "limpo". Este dicionário apresenta os nomes de localidades em minúsculas e sem quaisquer acentos, tal como o texto dos fólios. Assim, o processo consegue abranger um maior número de casos, o que torna a anotação automática mais eficaz e robusta.

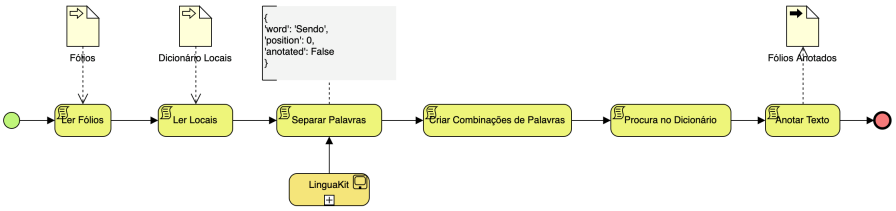


Figura 26: O processo de pesquisa com base em dicionário de locais

Como se pode observar na Figura 26, este processo, tal como o anterior, inicia a sua atividade com a leitura dos textos dos fólios (Ler Folios) e, recorrendo ao dicionário de locais criado anteriormente (Ler Locais), que tem os locais armazenados numa estrutura de dados, faz a separação dos texto por palavras (Separar Palavras), de modo que o módulo consiga comparar os locais do dicionário e os locais contidos no texto. Tal como no processo anterior de identificação de potenciais locais, neste processo também foi utilizada a ferramenta *LinguaKit* (CILENIS SL, 2020). No entanto, desta vez não foi utilizada a funcionalidade que faz a partição do texto e o armazenamento das palavras do documento numa estrutura de dados (Separar Palavras). Nesta estrutura de dados, para além das palavras encontradas é ainda armazenada a sua posição, juntamente com um valor (verdadeiro ou falso) que informa se essa palavra já foi anotada ou não. Inicialmente este valor assume o valor de falso. Terminada esta tarefa, faz-se a combinação de palavras (Criar Combinação de Palavras). Esta tarefa precisa de ser realizada porque existem nomes de localidades constituídas por mais do

que uma palavra. Como as localidades portuguesas não apresentam, usualmente, mais do que cinco palavras juntas, definimos um máximo de cinco palavras por combinação. Tendo as palavras separadas e as combinações realizadas, o processo compara as palavras contidas no texto com as referências de locais que estão registadas no dicionário (Procura no Dicionário), com o objetivo de as classificar, de facto, como locais. Como já referido, para evitar erros na escrita dos locais nos documentos, o processo de comparação utiliza palavras em minúsculas e sem acentos, usando assim o dicionário de locais "limpo". Com esta estratégia conseguimos cobrir um maior número de casos, tornando a anotação automática mais eficaz e robusta. Por fim, faz-se a anotação dos locais encontrados (Anotar Texto).

5.4 A integração do módulo de georreferenciação

5.4.1 *O sistema de Backend*

Na implementação da parte de *backend* do sistema Tommi utilizámos o *Flask* (Pallets, 2021), que é uma *framework* de desenvolvimento Web para *Python*. Dado que esta *framework* foi desenvolvida em *Python*, a integração dos módulos de georreferenciação que foram desenvolvidos foi realizada de forma bastante linear, não tendo apresentado grandes dificuldades na sua concretização. Todavia, a sua integração na estrutura do sistema impôs o estabelecimento de novas rotas necessárias para estabelecer a ligação com as aplicações de processamento dos textos, bem como para a obteção dos dados resultantes desse processamento. Os dados resultantes do processamento são armazenados na base de dados do sistema e, posteriormente, apresentados nos sistemas de interface desenvolvidos especificamente para o efeito.

De modo a satisfazer as necessidades da interface gráfica criada para os módulos de georreferenciação, fizemos a criação de três novas rotas no servidor de backend, nomeadamente a de:

- processamento dos dados, ou seja, a rota responsável pela anotação dos locais que aparecem nos documentos inseridos no sistema *Tommi* e pelo seu armazenamento na base de dados, tanto em formato regular como em formato anotadado;
- obtenção de todas as localidades encontradas que estão armazenadas na base de dados, realizando a filtragem de todas aquelas que não foram removidas parcialmente;
- execução de remoções parciais, ou seja, pela procura da localidade que se pretende remover; esta rota é bastante útil, para que, na fase de processamento de dados, a localidade não volte a ser inserida na base de dados, o que permite ao nosso sistema saber que essa localidade ou não está presente no documento, ou então está errada a sua anotação.

Com as rotas referidas criadas e os módulos de georreferenciação implementados e colocados no servidor, a integração com o servidor de *backend* ficou completa. Desta forma, o

sistema incorporou os serviços necessários para fazer a anotação dos locais referidos no Livro das Propriedades.

5.4.2 *O sistema de FrontEnd*

Após a modificação do sistema de *backend* passámos à modificação do sistema de *frontend*, de forma a incorporar os diversos componentes responsáveis pela invocação dos serviços de georreferenciação que foram implementados. Neste processo utilizámos a *VueJs* (Vue, 2021), uma *progressive framework* em *JavaScript* que nos permite de forma muito versátil construir sistemas de interface bastante sofisticados. Na criação da parte do *frontend* relativa aos serviços de georreferenciação foram projetadas e implementadas três *views*, isto é, três páginas especialmente concebidas para suportarem esta parte do sistema de interface. Na Figura 27, podemos observar as *tags* que envolvem o código *HTML* da página *web* em *VueJS*.

```
<template>
// Uso dos componentes importados na forma de tags HTML
</template>
```

Figura 27: Esqueleto de uma View em VueJS

Em *VueJS* existe uma separação por *tags* do código *HTML* do código *JavaScript*. É separada a parte que desenha o nosso *component*, por exemplo uma tabela na aplicação *web*. Da parte que invoca os *components* a desenhar se obtêm os dados que vão fazer parte, por exemplo, da tabela desenhada. As *views* como se pode observar na Figura 28, servem para invocar os *components*, que são peças mais pequenas da nossa aplicação, como uma tabela, e servem ainda para organizar a forma como os *components* são desenhados na aplicação *web*, criando assim uma página *web*, completa.

```
<script>
// Importação de pacotes e ficheiros necessários tais como os componentes
export default {
    components:{
        // Dar nome a todos os componentes importados
    },
    created() {
      // Faz o pedido ao servidor de backend no momento de criação desta página
    }
}
</script>
```

Figura 28: Esqueleto do Script de uma View/Component

Para que seja possível a navegação entre as páginas da aplicação *web* foi necessária a criação de um *Router*, que tem como função indicar que componente será desenhado na aplicação *web* quando entramos numa determinada rota (*url*) no nosso *browser*, e.g Figura 29.

```
{
    path: '/home',
    name: 'home',
    component: () => import('../views/Home.vue')
}
```

Figura 29: Exemplo de uma rota definida em Vue

Em suma, para desenvolver a parte do sistema de interface para os serviços de georreferenciação, foram desenvolvidas quatro *views*. De referir, uma para criação de uma listagem em formato tabular de todos os locais identificados, uma para realizar o pedido de processamento dos textos contidos no sistema, e duas que apresentam a integração com o *google maps*, georreferenciando todos os locais ou um local conforme a intenção do utilizador. Estas *views* necessitaram ainda da criação de quatro *components* para desenhar na interface tudo o que era pretendido.

Para usar o módulo de georreferenciação o utilizador precisa de aceder à barra de navegação do sistema, em particular ao menu de georreferenciação. Neste menu estão disponíveis três opções: processamento, mapas e localidades. A primeira destas opções permite aceder a uma aplicação de consulta, que informa o utilizador sobre o processamento dos textos à medida que ele se vai realizando. No fim deste processamento a anotação das localidades referidas nos textos foi realizada e o seu resultado inserido na base de dados do sistema. No final deste processo o sistema redireciona-nos para a página da aplicação da gestão de locais (Figura 30)

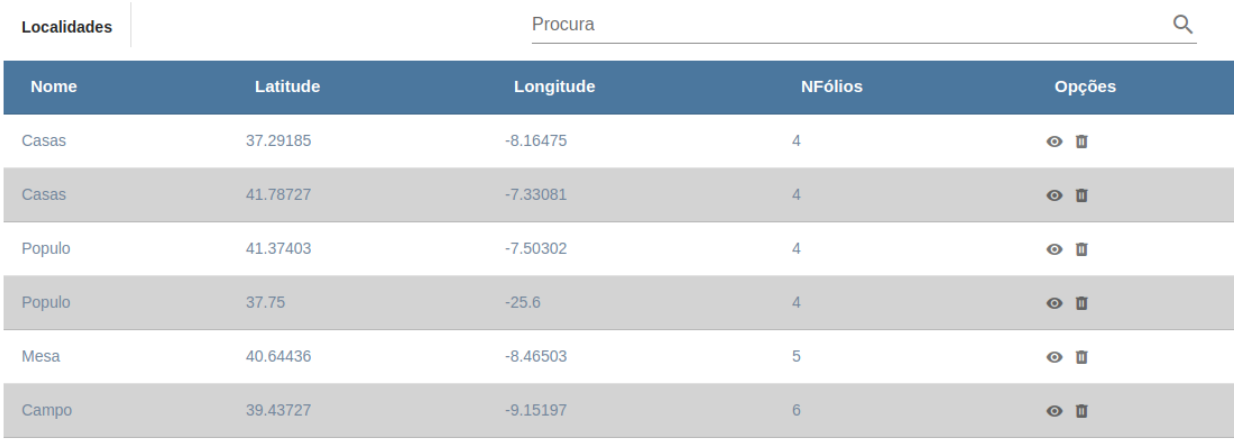


Figura 30: Fragmento da lista de locais anotados no sistema

Para que possamos entender de forma mais clara a forma como o sistema de georreferenciação atua, facilitando um pouco mais a utilização do sistema implementado, desenvolvemos um pequeno diagrama de atividades Figura 31, que resume o conjunto de tarefas que o sistema realiza no âmbito do processo de georreferenciação de locais do Livro das Propriedades.

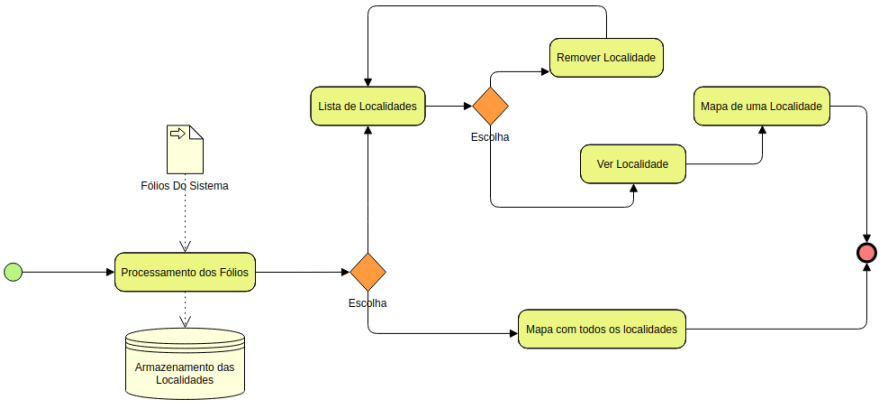


Figura 31: O processo de Georreferenciação do sistema *Tommi*

Como se pode ver através da Figura 31, o início do processo de georreferenciação começa com o processamento dos textos (Processamento dos Fólios) já inseridos na base de dados do sistema. Tal como referido anteriormente, como resultado deste processamento são armazenados no sistema os diversos locais que foram encontrados nos textos analisados, acompanhados pelas suas coordenadas geográficas e uma relação dos fólios nos quais foram encontrados. Com o processamento realizado e os locais devidamente anotados, podemos pode optar por visualizar os seus dados num mapa. Na lista de locais que o sistema nos apresenta podemos observar as localidades que foram anotadas pelo sistema, bem como as suas coordenadas geográficas e os números de fólios nos quais essa localidade é referida. Através desta lista, selecionando o ícone de visualização, o utilizador é redirecionado para a zona de visualização das topónimos num mapa (Figura 32)

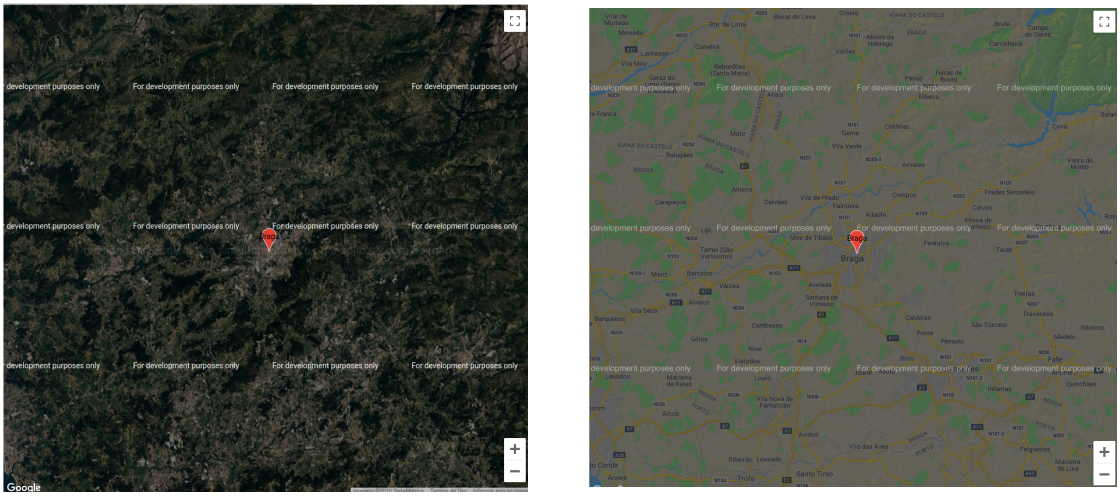


Figura 32: Mapa com representação de uma localidade, em modo satélite e estrada

O sistema de visualização de mapas que foi integrado no sistema oferece três tipos de visualização de dados diferentes, nomeadamente: uma vista de satélite; uma vista de estrada e uma vista híbrida, que é uma combinação das duas vistas anteriores. Estes tipos de vistas podem ser selecionados através de um conjunto de botões (Figura 33) que aparece por cima

das imagens dos mapas. Com estas opções é possível alterar a forma como se vê o mapa, de forma que o utilizador possa ter uma perceção diferente da posição na qual o local foi anotado.

Figura 33: Opções de mudança de Forma de Mapa

Além disso, podemos ver todas as localidades que foram identificadas de forma georreferenciada no mapa (Figura 34). Desta forma, através do sistema e dos novos serviços implementados, o utilizador consegue usufruir de um sistema de georreferenciação de texto, que lhe permite consultar os locais que foram identificados nos textos do Livro das Propriedades, bem com percorrer a zona assinalada, observando o que rodeia esse mesmo ponto geográfico.

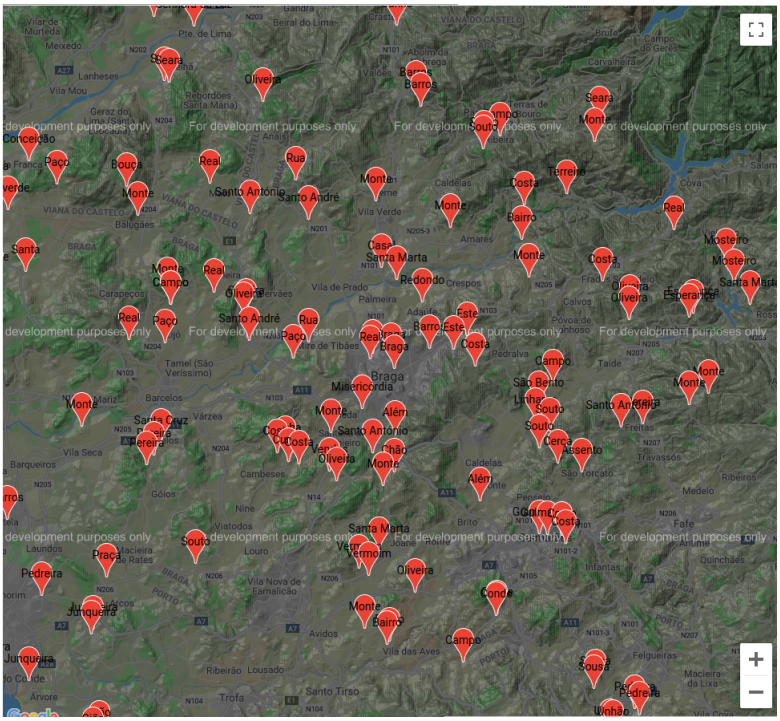


Figura 34: Vista de todas as localidades georreferenciadas

Por último temos a funcionalidade ao clicar na marca de sinalização no mapa e obter informações tais como o número de fólios em que esse nome de lugar aparece e informações geográficas acerca da topónimo.

6

CONCLUSÕES E TRABALHO FUTURO

6.1 CONCLUSÕES

Neste trabalho de dissertação concebemos e desenvolvemos um módulo de georreferenciação de texto para o sistema *Tommi* Barros et al. (2020), com o objetivo de identificar e situar geograficamente os locais referidos nos textos do *Livro das Propriedades* (Barros, 2019) (Barros, 2021). Com este módulo, os utilizadores e estudiosos do manuscrito, quer estes sejam professores, investigadores ou estudantes, poderão adquirir conhecimento complementar acerca da área, local e propriedades que estão referidos nos inúmeros registos de propriedades que figuram no Livro das Propriedades. A criação deste módulo implicou a implementação de um processo específico de anotação de referências de locais contidas nos textos, recorrendo a técnicas de PLN, bem como de um processo de associação de dados georreferenciados implicitamente a dados georreferenciados explicitamente. Estes dois processos juntos deram origem àquilo que designamos por *módulo de georreferenciação de texto*. O módulo de georreferenciação foi desenvolvido ao longo de várias etapas. Começámos por definir aquilo que pretendíamos para o sistema, determinando as suas principais funcionalidades e os seus métodos de tratamento de dados. Com isto, estabelecemos, depois, os dicionários de locais (*gazetters*), com as coordenadas geográficas, concebemos o processo de *geoparsing*, e incorporámos os dados georreferenciados num mapa, realizando assim o processo de *geocoding*. Antes e começar a anotar os locais, com base no dicionário de locais criado, implementámos um módulo de identificação de potenciais locais, ou seja, um componente que permitisse identificar locais referidos nos textos que não constassem no dicionário de locais criado. Com este módulo foi possível encontrar locais que hoje têm designações diferentes daquelas que foram utilizadas nos textos. Na implementação deste módulo utilizámos a ferramenta *Linguakit* (CILENIS SL, 2020), que nos permite classificar palavras a partir do texto. Após a implementação e validação destes módulos fizemos a sua incorporação no *backend* do sistema *Tommi* e desenvolvemos um sistema de interface que fosse capaz de disponibilizar aos utilizadores do sistema uma forma fácil e expedita de lidar com os módulos de georreferenciação desenvolvidos, em particular fornecendo meios para processar os textos do *Livro das Propriedades* que estivessem inseridos no sistema, georreferenciar os seus topónimos, consul-

tar as localidades e demais pontos geográficos identificados e a sua representação num mapa, entre outras funcionalidades.

Ao longo da implementação deste sistema, surgiram algumas dificuldades, em particular no tratamento de dados do *GeoNames*, visto estarmos a tratar com um número elevado de dados, com muita informação armazenada, tornando difícil filtrar a informação pretendida e processar os dados de forma a obter o dicionário de locais pretendido. Além disso, foi difícil perceber se os dados disponíveis estavam corretos e as coordenadas geográficas associadas bem atribuídas, e ainda estabelecer uma forma de anotação efetiva, que considerasse a existência de casos de paragem e identificadores de classe a partir dos documentos disponíveis. Por fim, enfrentámos alguns obstáculos de relevo durante o processo de associação das coordenadas dos locais a um mapa, que emergiram por causa dos serviços de mapas disponíveis apresentarem limitações, próprias de uma versão gratuita. Neste trabalho utilizámos a versão gratuita e de livre acesso do *Google Maps* (Google, 2021).

Compensando um pouco estas dificuldades, realce-se a ferramenta *Linguakit*. Esta ferramenta possui várias funcionalidades compatíveis com a língua portuguesa, o que permitiu acelerar o processo de implementação do sistema de georreferenciação. Além disso, apesar de terem implicado algum trabalho de preparação, os dados do *geonames* (Kaggle, 2020) foram muito úteis dada a sua atualidade, permitindo o acesso aos locais atuais de Portugal e, consequentemente, facilitando o trabalho realizado. Terminamos esta secção com uma última referência positiva à integração do sistema que implementámos com o *Google Maps*. Embora o seu processo tenha tido algumas contrariedades, como referido, esta ferramenta permitiu uma integração muito fácil com o sistema de interface que criámos, oferecendo funcionalidades que facilmente nos permitem associar as coordenadas geográficas de locais a um mapa.

Ao longo desta dissertação foram superadas várias dificuldades, encontradas quer ao nível da implementação do sistema, quer ao nível das decisões que tiveram que ser tomadas. Gradualmente, estas dificuldades foram ultrapassadas e o processo de desenvolvimento foi avançando, produzindo, por fim, um sistema de georreferenciação provido de um conjunto de funcionalidades importante, que nos permite ter uma melhor "visão" dos locais referidos no *Livro das Propriedades*, algo essencial na ajuda a quaisquer utilizadores, investigadores, professores e alunos que pretendam explorar as localizações geográficas dos bens inventariados no *Livro das Propriedades*, num contexto didático, não só para o estudo da língua portuguesa, mas sobretudo para o conhecimento geográfico e humano do norte de Portugal, além de outras áreas, na época na qual o livro foi redigido.

6.2 Trabalho Futuro

Apesar de o sistema de georreferenciação ter sido realizado e integrado no sistema de forma efetiva, reconhecemos nesta altura que alguns dos seus serviços e funcionalidades poderão ser aprimoradas, de modo a tornar o sistema criado mais robusto e melhorar a experiência de

utilização dos seus utilizadores. Por exemplo, poderíamos conceber e implementar um módulo de aprendizagem supervisionada no sistema para melhorar a precisão de anotação de locais no sistema e aperfeiçoar a sua georreferenciação. Este tipo de abordagem, quando devidamente implementado, poderá reduzir significativamente o erro na anotação dos locais e, consequentemente, melhorar os serviços de georreferenciação agora criados. O módulo de identificação de potenciais locais poderá ser também expandido com novos identificadores de classe para aumentar o número de locais anotados que não constem no dicionário de locais, podendo ainda aperfeiçoar-se a anotação de potenciais locais através do desenvolvimento de um serviço específico para a descoberta das coordenadas geográficas desses locais, que utilize, por exemplo, *Web Scraping*. Com tempo e recursos adequados, todas estas funcionalidades podem ser implementadas, o que oferecerá ao sistema Tommi um conjunto de serviços de anotação de locais mais preciso e robusto.

BIBLIOGRAFIA

Vue.js, 2021. URL https://vuejs.org/.

Beatrice Alex, Kate Byrne, Claire Grover, and Richard Tobin. Adapting the Edinburgh Geoparser for Historical Georeferencing. *International Journal of Humanities and Arts Computing*, 9(1):15–35, mar 2015. ISSN 1753-8548. doi: 10.3366/ijhac.2015.0136.

Anabela Barros, Orlando Belo, João Gomes, Tiago Fraga, Ricardo Martins, and José Pedro Carvalho. A Computational Instrument for students accessing and exploring the book of properties of the Braga archbishop's table (17th Century). pages 1–7, 2020.

Anabela Leal Barros. Apontamentos lexicais sobre o Livro das Propriedades ou Tombo da Mitra Arquiepiscopal de Braga: designações de terras e outros aspectos das propriedades. In *Estudos da linguística histórica: mudança e estandardização*, pages 393–428. Coimbra: Imprensa da Universidade de Coimbra, 2019.

Anabela Leal Barros. A edição do Livro das Propriedades ou Tombo da Mitra Arquiepiscopal de Braga. In *Paulo Abreu et alii, Os sete castelos. Congresso de Homenagem a D. Rodrigo de Moura Teles, Braga(no prelo)*, 2021.

Jesús Cascón-Katchadourian, Antonio Ángel Ruiz-Rodríguez, and Jordi Alberich-Pascual. Uses and applications of georeferencing and geolocation in old cartographic and photographic document management. *Profesional de la Informacion*, 27(1):202–212, 2018. ISSN 16992407. doi: 10.3145/epi.2018.ene.19. URL https://doi.org/10.3145/epi.2018.ene.19.

Hao Chen, Maria Vasardani, and Stephan Winter. Georeferencing places from collective human descriptions using place graphs. *Journal of Spatial Information Science*, 17(17):31–62, 2018. ISSN 1948660X. doi: 10.5311/JOSIS.2018.17.417.

CILENIS SL. Análisis completo, 2020. URL https://linguakit.com/es/analisis-completo.

Duarte Dias, Ivo Anastácio, and Bruno Martins. A Language Modeling Approach for Georeferencing Textual Documents. *II Congreso Español de Recuperación de Información CERI 2012*, 2009:85–96, 2012.

Patrick Drouin. Review of Jackson & Moulinier (2002): Natural language processing for online applications: Text retrieval, extraction and categorization. *Terminology. International*

Journal of Theoretical and Applied Issues in Specialized CommunicationTerminology / International Journal of Theoretical and Applied Issues in Specialized CommunicationTerminology, 10(1): 177–179, 2004. ISSN 0929-9971. doi: 10.1075/term.10.1.12dro.

Stuart E Dunn. Georeferencing: The Geographic Associations of Information.Hill, Linda L. *Literary and Linguistic Computing*, 22(3):367–369, 2007. ISSN 0268-1145. doi: 10.1093/llc/fqm014. URL https://doi.org/10.1093/llc/fqm014.

ESRI. ArcGIS - Imagery (WGS84), 2020a. URL https://www.arcgis.com/home/webmap/viewer.html?webmap=52bdc7ab7fb044d98add148764eaa30a.

ESRI. ArcGIS Desktop | Documentation, 2020b. URL https://desktop.arcgis.com/en/.

ESRI. ArcGIS Online | Software de Mapeamento SIG Baseado na Cloud, 2020c. URL https://www.esri.com/pt-pt/arcgis/products/arcgis-online/overview.

Judith Gelernter and Shilpa Balaji. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667, oct 2013. ISSN 13846175. doi: 10.1007/s10707-012-0173-8.

Google. Overview | Maps JavaScript API | Google Developers, 2021. URL https://developers.google.com/maps/documentation/javascript/overview.

Ian Gregory, Christopher Donaldson, Patricia Murrieta-Flores, and Paul Rayson. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9(1):1–14, 2015. ISSN 1753-8548. doi: 10.3366/ijhac.2015.0135.

Milan Gritta, Mohammad Taher Pilehvar, Nut Limsopatham, and Nigel Collier. What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623, 2018. ISSN 15728412. doi: 10.1007/s10579-017-9385-8.

Claire Grover, Richard Tobin, Kate Byrne, Matthew Woollard, James Reid, Stuart Dunn, and Julian Ball. Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1925):3875–3889, aug 2010. ISSN 1364503X. doi: 10.1098/rsta.2010.0149.

Kerric Harvey. Encyclopedia of Social Media and Politics. *Encyclopedia of Social Media and Politics*, (December), 2014. doi: 10.4135/9781452244723.

Infopédia. topónimo: Definição ou significado de topónimo no dicionário infopédia da língua portuguesa, 2021. URL https://www.infopedia.pt/dicionarios/lingua-portuguesa/topónimo.

Kaggle. GeoNames database, 2020. URL https://www.kaggle.com/geonames/geonames-database?select=geonames.csv.

Morteza Karimzadeh, Scott Pezanowski, Alan M. MacEachren, and Jan O. Wallgrün. GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, 23(1): 118–136, feb 2019. ISSN 14679671. doi: 10.1111/tgis.12510.

Jochen L. Leidner. Georeferencing: From Texts to Maps. In *International Encyclopedia of Geography: People, the Earth, Environment and Technology*, pages 1–10. John Wiley & Sons, Ltd, mar 2017. doi: 10.1002/9781118786352.wbieg0160.

Jochen Lothar Leidner. Toponym Resolution in Text Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Technical report, 2007.

Yolanda J. McDonald, Michael Schwind, Daniel W. Goldberg, Amanda Lampley, and Cosette M. Wheeler. An analysis of the process and results of manual geocode correction. *Geospatial Health*, 12(1), 2017. ISSN 19707096. doi: 10.4081/gh.2017.526.

Novetta. CLAVIN Documentation – Overview, 2019. URL https://clavin.bericotechnologies.com/clavin-core/.

Lluís Padró and Evgeny Stanilovsky. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, 2012. ELRA.

Pallets. Welcome to flask, 2021. URL https://flask.palletsprojects.com/en/2.0.x/.

Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fluart, Wajdi Zaghouani, Anna Widiger, Ann Charlotte Forslund, and Clive Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, (October):53–58, 2006.

Refinitiv. Intelligent Tagging & Text Analytics | Refinitiv, 2019. URL https://www.refinitiv.com/en/products/intelligent-tagging-text-analytics.

C. J. Rupp, Paul Rayson, Alistair Baron, Christopher Donaldson, Ian Gregory, Andrew Hardie, and Patricia Murrieta-Flores. Customising geoparsing and georeferencing for historical texts. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*, pages 59–62, 2013. doi: 10.1109/BigData.2013.6691671.

Franc Solina and Robert Ravnik. Georeferencing works of literature. *Proceedings of the International Conference on Information Technology Interfaces, ITI*, (May 2014):249–254, 2010. ISSN 13301012.

Geoffrey I. Webb, Johannes Fürnkranz, Johannes Fürnkranz, Johannes Fürnkranz, Geoffrey Hinton, Claude Sammut, Joerg Sander, Michail Vlachos, Yee Whye Teh, Ying Yang, Dunja Mladeni, Janez Brank, Marko Grobelnik, Ying Zhao, George Karypis, Susan Craw, Martin L. Puterman, and Jonathan Patrick. Density-Based Clustering. In *Encyclopedia of Machine*

Learning, pages 270–273. Springer US, 2011. doi: 10.1007/978-0-387-30164-8_211. URL https://link.springer.com/referenceworkentry/10.1007/978-0-387-30164-8{_}211.

Allison Gyle Woodru. Georeferenced Information Processing SYstem Christian Plaunt Step Two : Locating Pertinent Data. 1994.

Allison Gyle Woodruff and Christian Plaunt. GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655, 1994. ISSN 10974571. doi: 10.1002/(SICI)1097-4571(199410)45:9<645::AID-ASI2>3.0.CO;2-8.