APREGOAR: DEVELOPMENT OF A GEOSPATIAL DATABASE APPLIED TO LOCAL NEWS IN LISBON

Caroline Gilman Wentling

Dissertation presented as partial requirement for obtaining the degree of Master of Science in Geographical Information Systems and Science

NOVA Information Management School

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade NOVA de Lisboa

APREGOAR: DEVELOPMENT OF A GEOSPATIAL DATABASE APPLIED TO LOCAL NEWS IN LISBON

by

Caroline Gilman Wentling

Dissertation presented as partial requirement for obtaining the degree of Master of Science in Geographical Information Systems and Science

Adviser: Marco Octávio Trindade Painho

Co-adviser: Hugo Martins

November, 2022

STATEMENT OF ORIGINALITY

I declare that the work described in this document is my own and not from someone else. All the assistance I have received from other people is duly acknowledged and all the sources (published or unpublished) are referenced.

This work has not been previously evaluated or submitted to NOVA Information Management School or elsewhere.

NOVA Information Management School, November 22, 2022.

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.9.15) (Lourenço, 2021).

Acknowledgements

I would like to share my sincerest appreciation and thanks to all of those those who contributed to this work and my well-being while producing it.

To Professor Marco Painho for his guidance, feedback, and patience. For his instruction as I was just discovering GIS and its opportunities. Perhaps above all, for his confidence in me that transferred to confidence in myself.

To Professor Hugo Martins for the very long, technical phone calls of tool recommendations and project organizations, without which I might still be running in circles.

To Professor Miguel de Castro Neto and the team of NOVA Cidade, who opened new doors in the community for me and provided a working opportunity where I was able to develop skills that were directly translatable to this work. To Nuno Alpanhão who exposed me to new tools and resources and was always willing to help debug code. To Alexandre Baptista, who has eagerly assisted with translations and geo-annotation.

To Professora Sara Ribeiro, who provided many moments of emotional support and guidance throughout this process.

To Catarina Carvalho, whose enthusiasm about better connecting citizens to the news in their community reconfirmed my devotion to this project, and who contributed valuable information about current practices and future needs in the industry.

To Leon Sólon, for helping me learn how to learn to develop this kind of code.

To Lydia Schroder, with whom this concept was first sparked in Denver during long walking conversations of better communities in cities.

To the friends who have heard me speak about this project ad nauseam, for providing feedback or motivation to keep going. For incredible friendships and much needed distractions.

To the friends who have no idea what my thesis is about, yet have made me feel so glad to be who and where I am, and motivated to tackle new challenges in and out of my academic life.

To Bernardo Lago, for weathering my attempts build an application from scratch without any prior experience. For connecting me to people and resources to overcome

the many challenges I found myself facing. For his unyielding support, even when we found ourselves spending vacations coding until late into the night.

Of course, to my parents Caroline and Tom Wentling and brother Tommy for their unwavering support of and belief in me, for their hearing and feedback to countless versions of this concept, and for their their continued openness for this and all other adventures I've embarked upon thus far.

Also, to João Lourenço for creating the LaTex template (Lourenço, 2021) on top of which this documentation was developed.

APREGOAR: DESENVOLVIMENTO DE UMA BASE DE DADOS GEOESPACIAL APLICADA ÀS NOTÍCIAS LOCAIS EM LISBOA

Resumo

Há informações valiosas em formato de texto não estruturado sobre a localização, calendarização e a essências dos eventos disponíveis no conteúdo de notícias digitais. Vários trabalhos em curso já tentam extrair detalhes de eventos de fontes de notícias digitais, mas muitas vezes não com a nuance necssária para representar com precisão onde as coisas realmente acontecem. Alternativamente, os jornalistas poderiam associar manualmente atributos a eventos descritos nos seus artigos enquanto publicam, melhorando a exatidão e a confiança nestes atributos espaciais e temporais. Estes atributos poderiam então estar imediatamente disponíveis para avaliar a cobertura temática, temporal e espacial do conteúdo de uma agência, bem como melhorar a experiência do utilizador na exploração do conteúdo, fornecendo dimensões adicionais que podem ser filtradas.

Embora a tecnologia de atribuição de dimensões geoespaciais e temporais para o emprego de aplicaçãoes voltadas para o consumidor não seja novidade, tem ainda de ser aplicada à escala das notícias. Além disso, a maioria dos sistemas existentes suporta apenas uma definição pontual da localização dos artigos, que pode não representar bem o(s) local(is) real(ais) dos eventos descritos.

Este trabalho define uma aplicação web de código aberto e uma base de dados espacial subjacente que suporta i) a associação de múltiplos polígonos a representar o local onde cada evento ocorre, os prazos associados aos eventos, em linha com os atributos temáticos tradicionais associados aos artigos de notícias; ii) a contextualização de cada artigo através da adição de mapas de eventos em linha para esclarecer aos leitores onde os eventos do artigo ocorrem; e iii) a exploração dos corpora adicionados através de filtros temáticos, espaciais e temporais que exibem os resultados em mapas de cobertura interactivos e listas de artigos e eventos.

O projeto foi aplicado na área da grande Lisboa de Portugal. Para além da funcionalidade acima referida, este projeto constroi gazetteers progressivos que podem ser reutilizados como associações de lugares, ou para uma meta-análise mais aprofundada do lugar, tal como é percebido coloquialmente. Demonstra a facilidade com que estas dimensões adicionais podem ser incorporadas com grade confiança na precisão da

definição, geridas, e alavancadas para melhorar a gestão de conteúdo das agências noticiosas, a compreensão dos leitores, a exploração dos investigadores, ou extraídas para combinação com outros conjuntos dos dados para fornecer conhecimentos adicionais.

APREGOAR: DEVELOPMENT OF A GEOSPATIAL DATABASE APPLIED TO LOCAL NEWS IN LISBON

Abstract

There is valuable information in unstructured text format about the location, timing, and nature of events available in digital news content. Several ongoing efforts already attempt to extract event details from digital news sources, but often not with the nuance needed to accurately represent the where things actually happen. Alternatively, journalists could manually associate attributes to events described in their articles while publishing, improving accuracy and confidence in these spatial and temporal attributes. These attributes could then be immediately available for evaluating thematic, temporal, and spatial coverage of an agency's content, as well as improve the user experience of content exploration by providing additional dimensions that can be filtered.

Though the technology of assigning geospatial and temporal dimensions for the employ of consumer-facing applications is not novel, it has yet to be applied at scale to the news. Additionally, most existing systems only support a single point definition of article locations, which may not well represent the actual place(s) of events described within.

This work defines an open source web application and underlying spatial database that supports i) the association of multiple polygons representing where each event occurs, time frames associated with the events, inline with the traditional thematic attributes associated with news articles; ii) the contextualization of each article via the addition of inline event maps to clarify to readers where the events of the article occur; and iii) the exploration of the added corpora via thematic, spatial, and temporal filters that display results in interactive coverage maps and lists of articles and events.

The project was applied to the greater Lisbon area of Portugal. In addition to the above functionality, this project builds progressive gazetteers that can be reused as place associations, or for further meta analysis of place as it is colloquially understood. It demonstrates the ease of which these additional dimensions may be incorporated with a high confidence in definition accuracy, managed, and leveraged to improve news agency content management, reader understanding, researcher exploration, or extracted for combination with other datasets to provide additional insights.

Palavras-chave

noticiários geo-referenciados

gazetteer progressivo

geo-anotação

neogeografia

aplicação de web

informação geográfica voluntária (VGI)

Keywords

geo-referenced news media

progressive gazetteer

geo-annotation

neogeography

web application

volunteered geographic information (VGI)

Glossary

accuracy "how well the marker approaches the true spatial dimensions of the attribute being mapped" (Brown & Pullar, 2012) (pp. 7, 16, 48, 58, 77)

actively shared data collected that has been consciously initiated by a participant (pp. xvi, 10, 72)

appellation formation the removing the designator of a specific thing (Al-Olimat et al., 2018) (p. 73)

application programming interface a defined interface that allows different computer programs to communicate (pp. xvii, 8)

buffer zone "the area within a specified distance to selected real world features" (QGIS project, 2022) (pp. 1, 76)

Carta Administrativa Oficial de Portugal administrative and/or statistical limits and associated information of Portugal, corresponding to the NUTS classifications (pp. xvii, 29)

category ellipsis the stripping of "words related to the location category" (Al-Olimat et al., 2018) (p. 73)

comma separated value text file of data records (features) in which each record is stored as a new line and its attributes (fields) are delimited by a comma (pp. xvii, 58)

Creative Commons a non profit that provides, among other things, access and permissions for anyone or organization to use, rework, or redistribute with appropriate attribution to the original provider (Creative Commons, n.d.) (pp. xvii, 32)

Câmara Municipal municipal chamber; the executive body of each Portuguese municipality and, by extension, its departments and services (p. xvii)

endonym "The locale name for an object within a language area" (Witschas, 2004); "a locally used toponym" (Nordquist, 2018) (pp. 13, 83)

envelope minimum bounding box of a geometry, usually defined by the coordinates of its four corners (pp. 30, 33)

event-irrelevant those locations occur "when the raw texts contain news summaries of events that are not of interest" (Lee et al., 2019) (p. 83)

event-occurring "all locations where events occurred regardless of whether the event is the event of interest" (Lee et al., 2019) (p. 83)

event-relevant "those locations that are part of the main description of the event of interest, i.e., all locations that are key to the narrative of the event of interest" (Lee et al., 2019) (pp. 78, 83)

exonym "a place name that isn't used by the people who live in that place but that is used by others" (Nordquist, 2018) (p. 13)

explicit metonymy the removing of the thing and leaving only the designator (Al-Olimat et al., 2018) (p. 73)

Extensible Markup Language a computer and human readable markup language for the transfer and storage of data (p. xix)

free software "software that respects users' freedom and community. Roughly, it means that the users have the freedom to run, copy, distribute, study, change and improve the software" (Free Software Foundation, 2022) (pp. xvii, 55)

freguesia the Portuguese word for the equivalent of civil parish, indicating subdivisions of municipalities, and the smallest administrative division in Portugal (pp. xii, 21, 23–25, 28, 29, 43, 45, 51, 52, 54, 74)

gazetteer "a geographical dictionary, most commonly containing place names and associated properties such as geographic coordinates, type of place, and population, among others" (Karimzadeh et al., 2019) (pp. 2, 13–17, 21, 23, 28, 30, 37, 46, 60, 73, 76–79)

geo-annotation "the process of manually tagging (segmentation) and annotating place names in text with entries (toponyms) from a geographic gazetteer" (Karimzadeh et al., 2019) (pp. 2, 13, 16, 17, 19, 33, 79)

geocoding ”the process of taking input text, such as an address or the name of a place, and returning a latitude/longitude on the Earth’s surface for that place”(Gupta & Nishu, 2020) (pp. 13, 15, 17, 79)

geocorpora news articles of a variety of different publication sources introduced into the Apregoar geodatabase that serve as the test data for the Apregoar system (pp. 24–26, 28, 33, 36, 37, 46, 48, 50, 54, 59)

geographic information system a framework for the manipulation and analysis of geographic data (p. xvii)

Geography Markup Language an open standard for data exchange defined by the OGC, which extends XML to express geographical features (pp. xvii, 68)

geolocation the definition of a point in space relative to the earth’s surface (pp. xii, 5, 7, 12, 16, 17, 69, 76, 79)

geoparsing “the process of automatically resolving place reference in natural language (unstructured text) to toponyms in a geographic gazetteer with geographic coordinates” (Karimzadeh & MacEachren, 2019) (pp. 13, 14, 16, 73, 74)

geoportal ”the human-to-machine interface performing as a single point-of-access to spatial data and geo-information systems, offering sharing capabilities and connecting between geospatial data providers and end users,... typically employed as a web-based GUI equipped with functionalities for accessing Earth observation data and geographical information”(Jiang et al., 2020) (pp. 7, 8, 55)

georeferencing the association of ”locations or location-related data content onto a map” (Xing et al., 2015) (pp. 6, 19, 48)

geospatial intelligence the ”exploitation and analysis of satellite imagery and other forms of earth observation data to describe, assess, and visually depict physical features and geographically referenced activities on the earth” (Datta, 2018) (pp. xvii, 10)

geotag ”the process of identifying and disambiguating references to geographic locations (i.e., toponyms)... consists of two steps: toponym recognition, where all toponyms (e.g., “Paris”) are identified, and toponym resolution, where each toponym is assigned to the correct geographic coordinates among the many possible interpretations (e.g., “Paris” which can be one of over 140 places including France and also Texas).” (Lieberman et al., 2010) (pp. 13, 30, 33, 73, 78–80)

global positioning service a network of earth orbiting satellites and reception devices that are used to determine geolocation of objects (pp. xvii, 5)

graphic user interface in this work, the visualization of the response of an application request via the Flask view functions and methods for users to interact with the application (p. xviii)

ground truth a hand coded set of locations considered true for training and verification(Lee et al., 2019), however these are subject to annotator error or misunderstanding and may not be exactly where events occurred (pp. 11, 16, 17, 20)

hyperlocality “a spatiality that is endemic - i.e., locationally specific - to the individual, real-time positionalities of digital platform users” (Leszczynski, 2019) (pp. 7, 69)

informative corpora news stories of a variety of different publication sources from which relevant insights and attributes were extracted to inform the data model and ultimately the various UIs of the web app (pp. xxiii, 26, 81)

internet of things the integration of networked hard- ware sensors to monitor and/or interact with their surroundings (pp. xviii, 53)

Javascript object notation a format for data-interchange built on name/value pairs (objects) and ordered lists of values (arrays) that is straightforward for humans and computers to generate and parse (ECMA International, 2017) (pp. xviii, 30)

Jinja A templating tool by Pallet Projects that supports a Python-like syntax of code to include passed variables before being rendered in the final document (Pallets, n.d.-b) (pp. 32, 33)

Junta de Freguesia parish council, organizing body of a freguesia (pp. xviii, 25)

location ellipsis — the dropping of the specific location reference in the location name (Al-Olimat et al., 2018) (p. 73)

mashup — "a Web application that aggregates multiple services to achieve a new purpose" (Xing et al., 2015) (p. 8)

materialized view — saved and named complex queries within the PostgreSQL vocabulary, the results of which are persisted in table-like form resulting in faster recall in comparison to the view, though requiring an intentional refresh to update its contents (p. xvi)

monitor — in the context of this work, anyone leveraging local data to improve the community, such as city officials, city planners, local activists, etc. (pp. 2, 3)

named entity recognition — "the identifying entities such as person, location, and organization names" (Teitler et al., 2008); "the task of extracting and distinguishing different types of entities in text (i.e. names of people or organizations, dates and times, events, geographic features or even 'non entities')" (Silva et al., 2006) (pp. xviii, 13)

namehead — "the complex phenomenon of alternate name forms" (Al-Olimat et al., 2018) (p. 73)

neogeography — access and use of geographic tools, creation of products, and contribution of geographic data by untrained users for personal or larger scale projects (Elwood et al., 2012) (pp. 7, 8, 50)

Nomenclature of Territorial Units for Statistics — "a hierarchical syster for dividing up the economic territory of the EU and the UK for the purpose of: the collection, development and harmonisation of European regional statistics, socio-economic analysis of the regions,... [and] framing of EU regional policies" (European Commission, n.d.) (p. xviii)

open data — data and content that "anyone can freely access, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)" (Open Knowledge Foundation, n.d.-b) (p. 68)

Open Database License a license managed by the Open Knowledge Foundation as a part of the Open Data Commons that permits anyone to share, create from, or adapt upon the database as long as that work properly attributes the original database, maintains the original licensing on the derived works, and provides access to unrestricted versions of the input database (Open Knowledge Foundation, n.d.-a) (pp. xviii, 14)

open source a development methodology, the product of which is free of any restrictions of use, permits access to (for the study or modification of) the source code as well as the distribution of original or modified copies to third parties (pp. xviii, 55)

passively shared data collected in the background of regular activity (pp. xvi, 72)

place a locality associated with communal and/or individual social understanding(s) (pp. 1, 2, 5–7, 12–21, 27–29, 34, 37, 46, 48, 52, 72–76, 79, 83)

point of interest any entity (natural or artificial) with a well-defined location; ex: Praça do Comércio or Garden of the Gods (pp. xviii, 7, 15)

precision "a measure of the exactness in placing feature marker on the map", especially relevant at smaller scales (Brown & Pullar, 2012) (pp. 7, 16, 58)

proof of concept functional or demonstrative of the basic project concepts (Acedo et al., 2019) (pp. xviii, 3)

psychographics "the prevailing interests of people in an area" (Chiappinelli, 2020) (p. 71)

QGIS an open source GIS application supporting key vector and raster spatial manipulations and functionalities (QGIS Development Team, n.d.) (pp. 29, 30)

really simple syndication a standardized, computer-readable feed of updated content provided by a website such that it can be ingested by other, often third party, applications such as news aggregators (pp. xviii, 11)

reference ambiguity "the same location can have more than one name"(Silva et al., 2006) (pp. 22, 52, 73)

referent ambiguity the same name can be used for more than one location"(Silva et al., 2006) (p. 73)

referent class ambiguity ”the same name can be used for locations as well as for other class of entities, like persons or company names”(Silva et al., 2006) (p. 73)

spatial attribute a description relating to location; ex: ‘where did something happen’ or ‘where was it logged’ (pp. xviii, 1)

spatial autocorrelation the phenomena that ”everything is related to everything else, but nearby things are more related than distant things”; the first law of geography (Tobler, 1970) (p. 23)

spatial data infrastructure an interconnected framework including geographic data, metadata and related tools, resources, and policies necessary to effectively manage, access, store, and use a variety of geodata for a broad range of purposes by various potential users (Jiang et al., 2020) (pp. xviii, 7)

spatial reference system a framework for establishing and comparing geolocations, comprised of a coordinate system for measuring location, a datum for translating these abstractions to earth space, and a projection for representing the 3D coordinates in a 2D plane (pp. xviii, xix, 5)

spatiality ”the way [one] interacts with space and other individuals on/in/through space” (Roche & Rajabifard, 2012) (pp. 18, 20)

temporal attribute a description of when; ex: ‘at what time did it happen?’ or ‘which day was it published?’ (pp. xix, 1)

thematic attribute a description of what, why, or how; ex: ‘what happened’ or ‘who published it’ (pp. xix, 1)

toponym a textual reference to geographic location (Lieberman et al., 2010) (pp. xii, xv, 13–16, 30, 43, 44, 73, 76–79, 83, 84)

toponym recognition the identification of place names in unstructured text (Halterman, 2019; Lieberman et al., 2010) (pp. xii, 13, 15)

toponym resolution the matching of a toponym to a set of geographic coordinates (Halterman, 2019; Lieberman et al., 2010) (pp. xii, 13, 15)

union the combination of two datasets into a single result (PostgreSQL Tutorial, 2022) (pp. xxiv, 21, 22)

user interface the method of interaction between a user and the program (p. xix)

view saved and named complex queries within the PostgreSQL vocabulary that are run every time they are referenced, resulting in updated yet potentially slower results in comparison to the materialized view (PostgreSQL Tutorial, 2022) (pp. xiii, 47)

view function the code responding to requests to the Flask application, capable of processing and returning data, rendering templates, and/or redirecting as necessary (pp. xii, 33)

volunteered geographic information geographic user generated content, which encompasses any and all human geo-annotated data regardless of expertise, and enabled by internet access, positioning devices, and applications facilitating actively shared or passively shared user input, supporting higher volumes of data generation and more personalized and transient products than are traditionally created by formal cartography (Elwood et al., 2012) (pp. xix, 7)

web app a program running on a web server that is accessible via a web browser with internet connectivity; short for 'web application' (pp. 3, 4, 18, 25–27, 30, 32–34, 36, 37, 47, 48, 50, 55, 58, 60, 86)

Werkzeug A WSGI web application library by Pallets Projects that supports communication with the Apache server, among other functions (Pallets, n.d.-c) (p. 31)

world geodetic system a geocentric system that "provides a basic reference frame and geometric figure for the earth, models the earth gravimetrically, and provides the means for relating positions on various geodetic datums and systems to an Earth-Centered, Earth-Fixed (ECEF) coordinate system" (DMA WGS 84 Development Committee, 1991) (pp. xix, 5)

Acronyms

AI	artificial intelligence (p. 79)
AM	A Mensagem (pp. 25–27)
AML	Área Metropolitana de Lisboa (pp. 21, 22, 28, 29, 74)
API	application programming interface (pp. 8, 20, 30, 44, 46, 49, 56–58, 68, 80)
ASA	Apregoar study area (pp. 21, 22)
B2C	business-to-consumer (p. 69)
CAOP	Carta Administrativa Oficial de Portugal (pp. 29, 46)
CC	Creative Commons (p. 32)
CML	Câmara Municipal de Lisboa (pp. 22–24, 26, 27, 29, 46, 49)
CRIL	Circular Regional Interior de Lisboa (p. 74)
CRUD	create, read, update, and delete (pp. 19, 68)
CSS	Cascading Style Sheets (p. 33)
CSV	comma separated value (p. 58)
DdL	Distrito de Lisboa (pp. 21, 22, 28, 29)
DNI Fund	Digital News Innovation Fund (p. 23)
DVL	data visualization literacy (p. 9)
FS	free software (p. 55)
GEOINT	geospatial intelligence (pp. 10–12, 14, 17, 52, 53, 57)
GeoJSON	geographic Javascript object notation (pp. 38, 56, 58, 68)
GIS	geographic information system (pp. xiv, 3, 6, 7, 19, 50, 52, 79)
GML	Geography Markup Language (p. 68)
GPS	global positioning service (pp. 5, 6, 29, 69)

GUI	graphic user interface (pp. xi, xxiv, 2, 3, 7, 19, 33, 38–42, 45, 47, 56, 59, 60, 86–90, 93)
HTML	HyperText Markup Language (p. 33)
HTTP	HyperText Transfer Protocol (pp. 30–32, 68)
ICT	information and communication technologies (pp. 9, 70)
ID	identifier (pp. 29, 39, 40, 43, 46)
IOT	internet of things (pp. 53, 70, 72)
JFC	Junta de Freguesia de Campolide (pp. 25–27)
JFCdO	Junta de Freguesia de Campo de Ourique (pp. 25–27)
JFE	Junta de Freguesia de Esrela (pp. 25, 26)
JS	JavaScript (pp. 30, 32, 33)
JSON	Javascript object notation (pp. 30, 33, 39, 40, 43, 45)
LBS	location based services (pp. 7, 69)
NER	named entity recognition (pp. 13, 76, 77)
NLP	natural language processing (pp. 16, 76, 78)
NUTS	Nomenclature of Territorial Units for Statistics (pp. ix, 21)
ODbL	Open Database License (pp. 14, 32)
OGC	Open Geospatial Consortium (pp. xi, 68)
ORM	object relational mapper (p. 32)
OS	open source (p. 55)
OSM	Open Street Map (pp. 29, 30, 44, 46, 48, 49, 79, 82)
OSMF	Open Street Map Foundation (p. 14)
POC	proof of concept (pp. 3, 50, 55)
POI	point of interest (pp. 15, 35, 48–50, 80, 84)
PPGIS	public participation geographic information system (pp. 52, 72)
RSS	really simple syndication (pp. 11, 77)
SA	spatial attribute (pp. 1–3, 18–20, 26)
SDI	spatial data infrastructure (pp. 7, 8, 58)
SQL	Structured Query Language (p. 32)
SRID	spatial reference system identifier (pp. 5, 29)

SRS	spatial reference system (pp. 5, 28, 68)
TA	temporal attribute (pp. 1–3, 18–20, 26, 69, 76, 78)
ThA	thematic attribute (pp. 1, 3, 6, 19, 20, 26, 34, 37, 69, 76)
UI	user interface (pp. xii, 77, 78)
UNGEGN	The United Nations Group of Experts on Geographical Names (p. 13)
URL	uniform resource locator (pp. 26, 31, 34, 41, 43, 79)
VGI	volunteered geographic information (pp. 7, 13, 14, 17, 30, 53, 72, 79)
WFS	Web Feature Service (pp. 32, 37, 39–41, 45, 47, 68)
WGS84	world geodetic system (p. 5)
WMS	Web Map Service (pp. 32, 39, 45, 47, 68)
WSGI	Web Server Gateway Interface (pp. xvi, 31, 32)
XML	Extensible Markup Language (pp. xi, 68)

Index of the Text

Index of Tables

Index of Figures

1

Introduction

1.1 Context

Our decisions are geospatial in nature (Bhattacharya & Painho, 2018). The value of spatial information goes beyond the technical and is already nestled into our everyday activities in the form of daily tasks such as navigation and service selection. Applications like Google Maps, AirBnB, and UberEats allow non-technical users to visualize and filter the distribution of various services through spatial attribute (SA), temporal attribute (TA), and thematic attribute (ThA). For example, a user on AirBnB may filter all apartments with high-speed Wi-Fi (ThA) available in the Estrela neighborhood and within walking distance to a market (SA) from Aug 1 to Aug 7, 2020 (TA).

Yet, though this type of manipulation is commonplace in the consumer products of many industries, it is glaringly absent from that of news media. When reading about an incident occurring in an unfamiliar place, readers will often need to look up the location. They may have trouble relating the spatial significance of an incident to neighboring occurrences or historical events in the same spot. Many articles define place via textual descriptions, but these can be easily overlooked if searched by keyword, especially if different names or alternate designations are employed by the searcher.

A place can be hard to pin down. Geographical cue words such as 'city of Lisbon', 'just outside of Lisbon', 'Lisbon-based', and 'Rio Tejo', (adapted from (Lieberman et al., 2010)) don't have precise spatial definitions, and can queue different understandings in various contexts. "Just outside of Lisbon" may not have a defined area, but it can situationally refer to a simple buffer zone of the city (such as "within 5km of the city") or a particular direction (Odivelas is just outside of the city, but other cities also border Lisbon in other directions). A reference to "Rio Tejo" should not necessarily associate a point to the geometrical center of the major river flowing across the country, but should consider the context of the story/place (the bank of Cais de Sodré, for example). In these cases, there is not an appropriate, automatic method to distil this from textual description. News articles also depend on nearby points of interest to situate their story, instead of using formal names. Though this may help to cue spatial contextualization

in readership, it depends heavily on readers already having an understanding of the spatial layout of an area (Lee et al., 2019).

This is a problem for researchers who may want to define a study area that does not conform to traditional administrative boundaries or existing points of interest, but also for the casual user or city official (monitor). The former might, while perusing headlines, miss an article of interest relating to a place along their commute home from work. The latter could be an elected official who seeks to monitor an issue (such as gentrification or homelessness) but is unable to visualize the subtle distribution of such events throughout his or her district. In such cases there is obvious disconnect between the existence of data and its usability. Though many search engine queries contain geographic keywords (Silva et al., 2006), news media enterprises have not yet accommodated such spatial associations to their articles that would provide an expected improved user experience and therefore competitive edge in their industry. As such, there is commercial as well as well as operational and academic value in better understanding the spatial distribution of events within a community, such that additional informative insights can be drawn.

1.2 Solution

This work seeks to explore the viability of journalists associating TAs and SAs to news stories, and discover if this data could provide additional and valuable information to a variety of users.

To accomplish this, a geo-annotation tool that allows journalists to define spatial and temporal dimensions to their news articles, and an exploration interface that allows users (journalist, readers, researchers, or monitors) to filter and extract news data of interest based on their SAs, TAs, and TAs is developed.

The following pages describe a set of functional tools that supports the creation and management of a spatial database of local news stories, a publishing interface (associating place to news events and adding these as records to the database), a variety of GUIs (search, filter and visualizations of results from the database), as well as a story contextualization feature (a map displaying the distribution of a story in line with its content). It also includes a progressive gazetteer of polygons, allowing journalists to assign either existing areas or define new areas, as most appropriate to the event at hand. This functionality should provide a basis from which meaningful projects may be developed for mass media applications in the future.

It is expected that the intentional association of place to traditional news articles will provide an added dimension of understanding to communities at a local level. This type of data preparation, though it is initially cumbersome to establish and requires adjustment of journalistic processes to maintain, will provide a powerful foundation from which future economic (improved publisher products elevating their offering and attracting/maintaining a customer base), societal (illumination of local trends requiring

intervention, improved community engagement of readers with their surroundings, or improved city resources), and academic (improved research functionality) benefits may stem. If this type of functionality and improved user experience are well-implemented by a handful of productive news services, perhaps it will inspire a shift of the industry standard towards integration of spatial attributes and spatially related products.

The project uses Lisbon, Portugal as a study area, applying multiple locally focused sources, including a local newspaper, one municipality administration, and two sub-municipality administrations to form a demonstrative corpora from which to apply, test, and demonstrate the functionality of the developed application. By building a tool specific to Lisbon, the project seeks to accommodate the culture and business processes of the local community, providing a platform that is useful and valuable to users (whether citizens, monitors, researchers, or news agencies).

Though testing of the hypothesis (manual, journalist association of SAs and TAs to the events covered in news articles will provide additional value to news agencies, readers, monitors, and researchers) through rigorous comparison to the status quo (traditional online news sources without a spatial element) and emerging product performing automatic extraction of place are not included in this endeavor, the resulting tools should provide a basis from which future projects may perform this evaluation.

1.3 Objectives

The proposed tangible results are a web app that allows non-technical users to explore spatial and temporal incident distributions within the chosen study areas. Its functionality includes:

1. A spatial database of incidents that supports the association of SAs, TA, and ThAs.
2. Publishing tool that allows journalists (or editors) to define the locations of events covered in a news story, also assigning TAs and traditional ThAs.
3. A contextualization tool that augments existing digital article webpages with maps showing the locations of the news events.
4. A searching interface for researchers that supports the filtering of SAs, TA, or ThAs in a familiar GUI, with results shared in list and map formats.

Beyond the implementation of this proof of concept (POC) toolset and demonstration of value, this project also seeks to tackle the following:

1. Planning and execution of a GIS product
2. Creation and maintenance of a geospatial database
3. Design and programming of user interfaces

4. Leveraging of open source programs and tools
5. Collaboration with news industry users
6. Development of a smart city product
7. Development of open source tools
8. Provisioning of the base product for future expansion into desired directions (integration of future language options, accommodation of multiple news sources, integration of planned events, integration of city resources, integration of automatically extracted place from historic sources, etc.)

1.4 Structure

This thesis is divided into six chapters. The first provides and overview of the Apregoar project concept and its objectives. The second chapter reviews existing literature on relevant spatial information, its effect on people, the parsing of spatial information from unstructured documents, and the opportunities and challenges of applying this to news media. The third chapter describes the methodology of the web app design, development, and testing. The fourth chapter is dedicated to the results and demonstration of spatio-temporally activated new articles in Lisbon, Portugal. The fifth chapter provides and analysis of the project and the limitations of the current system.The sixth and final chapter shares closing remarks.

2

Literature review

2.1 Spatial information

"Location is involved with everything" (Bhattacharya & Painho, 2018) and is even considered a key factor in decision-making (Roche & Rajabifard, 2012). In reference to a point relative to the earth's surface, geolocation (sometimes referred to herein as simply: location) and other attributes (thematic or temporal) describe phenomena such as where events happen or where something is spatially situated (Longley et al., 2005). While geolocation can be defined in a variety of ways, perhaps most commonly it is interpreted by humans in two dimensional coordinates associated with a particular spatial reference system (SRS). A common example is the often-used longitude and latitude coordinates of spatial reference system identifier (SRID) 4326, that is associated with a spheroidal cartographic reference surface of the world geodetic system (WGS84), and famously serving as the reference coordinates of the global positioning service (GPS) (DMA WGS 84 Development Committee, 1991). Such formal definitions of location allow the connection of disparate phenomena or datasets by providing a common framework within which they can be compared (Bhattacharya & Painho, 2018).

A place, by contrast to a location, is a geospatial definition, and potentially much more. A place includes the objective (descriptions of the objects physically present, such as a river or a building), but they also are subject to the multiple identities strewn upon them by those who experience them under different conditions, formed either by direct engagement or passed anecdotes (Massey, 1991). These dynamic and overlapping definitions mean that places can change over time and space. The same location can refer to multiple places, depending on the context. A place, then, can have history or nostalgia, and these informal relationships may be constantly morphing (Roche & Rajabifard, 2012). This is especially apparent when considering a microcosm such as a neighborhood, a community, or an area as a place. The experience of those who live, work, and visit a place constantly feeds back into these areas, using and affecting local knowledge and expectations of what the place is (Cai & Tian, 2016). For all of these

reasons, it can be challenging to associate ThAs attempting to characterize a place as it pertains to predefined boundaries, as the personal perception of these areas, even the locations of their borders, may differ from the official records (Acedo et al., 2019).

2.1.1 Geographic information systems

Though often relevant for decision making at the personal, organizational, and regional levels, spatial data is underutilized in formal analysis (Bhattacharya & Painho, 2018). GIS are computer programs that support the collection, sharing, processing and visualization of geospatial data and its resulting information (Baker et al., 2019). The association of data to a map (georeferencing) is fundamental for understanding where people, things, and events are, were, or may be (Rajabifard, 2009; Xing et al., 2015). The resulting geographical datasets, in the forms of maps and features, provide an opportunity to orient collaborators, share experiences, convey ideas, and challenge presumptions of the users (Baker et al., 2019; Jiang et al., 2020; Kleinhans et al., 2015). One of its most powerful opportunities of this dynamic description is to address problems by anchoring relations between datasets and developing spatially considerate solutions to identified problems (Bhattacharya & Painho, 2018; Rajabifard, 2009). It is no surprise, then, that location data is already considered valuable and increasingly being incorporated into at all scales of community operations (Bhattacharya & Painho, 2018; Roche & Rajabifard, 2012). One of the most valuable characteristics of GIS is the ability to display relational values across time and space (Baker et al., 2019). The convolution of the utilization of GIS technology, the availability of GPS infrastructure, and application of machine learning techniques promote a myriad of real-time and spatial services with research, commercial, and security implications (Al-Olimat et al., 2018; Barns, 2020).

2.1.2 Feature definition

Where something occurs is stored in a GIS in raster (pixelated image) or vector (feature) format. Vector features can range from simple geometries, such as points, lines, or polygons (0-, 1-, and 2-dimensional features, respectively), or develop into more complicated multi-element collections, mixed element collections, or three dimensional definitions (incorporating a z-axis in addition to x- and y-). The type of feature to be used for a particular application depends on the information one wants to convey. For example, points are presumed to have an unknown range of influence, whereas polygons impose a boundary on whatever they represent (Brown & Pullar, 2012). This can be problematic when attempting to describe a continuous value, or a feature with a fuzzy or inconsistent boundary. The choice also depends on the size of the dataset, as improved performance is noted in polygon representation of information of sparse sampling due to the high data density required for point analysis (Acedo et al., 2019).

The choice of feature type affect both the analytical processing as well as the deductions made when viewing the results, the latter being further affected by placement and marker representation (Brown & Pullar, 2012). Ultimately, the application and requirements of any spatial system will determine the selection of feature representation. Considering needs of accuracy versus precision, data literacy of the readers, opportunities for clarification from additional materials, and purpose will assist in the selection of feature type.

Cross-border mapping is a specific instance in which boundaries can be a challenge. Boundaries seem to instill a binary adherence, an inclusion or exclusion that, when applied to populations, can be analytically and socially divisive (Acedo et al., 2019; Massey, 1991). More likely, the data on either side of the line are heterogeneous, more fluidly changing state over a shore of values, rather than a counter-position to the opposite side of the enclosure (Witschas, 2004). In fact, political borders tend to be natural points of exchange of people, things, and ideas in a way that are least adherent to the administrative areas to which they belong (Xing et al., 2015). Massey even goes so far as to suggest that we can re-imagine place definitions as momentary states in time, or inextricably linked to the elements external to themselves (Massey, 1991). Therefore, any static representation is already a distortion of reality (just as, or more than, any data representation is an abstraction) (Baker et al., 2019).

One tool to address these fluid definitions is neogeography, a new way of understanding place by combining location based services (LBS) and volunteered geographic information (VGI) (Painho & Pina, 2013). These concepts may leverage hyperlocality, which incorporates more precise, potentially real-time locations (such as identification of actual user location on a map) into digital spatial tools (Imani et al., 2019; Leszczynski, 2019). As our movements have become greater in displacement and in number, so too have our connections to the people and places we encounter along the way. In that sense, communities are expanding beyond the confines of walkability and extending around the globe (Painho & Pina, 2013). Many of the same tools that allow us to remotely connect with each other (social media) include geolocation, making their users producers of vast quantities of facts and geographic references, the building blocks of geographic data (Longley et al., 2005). These can be used to identify new, digitally relevant point of interest locales where people convene, or be leveraged as expert data when users describe their own communities (Roche & Rajabifard, 2012).

2.1.3 SDIs and geoportals

A geoportal is a web-interface to a spatial data infrastructure (SDI), providing a user with the means to manipulate and view information drawn from the underlying datasets and GIS (Jiang et al., 2020). It presents information in the form of visualizations via a GUI, (Bhattacharya & Painho, 2018) which can be more intuitive and inviting to informal users than accessing information via building queries (see Appendix A.1 for

open geospatial standards and their application programming interface (API) services). Depending on the application, users may search and view datasets, leverage manipulation tools within the platform, and/or access the data for use outside the platform via an API or exports. The value of a geoportal is especially potent when incorporating a variety of distinct data sources, as the resulting system may integrate the disjoint data and augment the capabilities of the disparate systems (Bhattacharya & Painho, 2018). The resulting mashups allow users to combine existing data sources in new ways, (Xing et al., 2015), promoting access and sharing of new information (Jiang et al., 2020). A geoportal not only offers access of disparate datasets to end-users directly, but provides user-friendly tools that allow even non-technical users to extract value from the exposed products (Bhattacharya & Painho, 2018; Kleinhans et al., 2015), a foundational element of neogeography. This empowerment of the layperson to leverage their own content to make maps transforms the process of data collection, processing, and application to a less stringent, more democratic endeavor (Painho & Pina, 2013).

Though scientific research has historically devoured the most spatial data (Bhattacharya & Painho, 2018), the use of geoportals has expanded to include international organizations, governmental agencies, and commercial purposes as key industries (Jiang et al., 2020). Governments, for example, have recently been recognizing the opportunities of open government policies and of geoportals to support the access, openness, transparency, and accountability inherent in these kinds of policies (Jiang et al., 2020). Meanwhile, the commercial sector may be driving the availability of different kinds of datasets by recognizing the power of heterogeneous data sources in their own analysis or products, and use their own resources to contribute to the available data pools or tool sets by which to manipulate those data (Afzalan et al., 2017; Jiang et al., 2020). This kind of collaboration between usually disparate systems facilitates the application of new technological methods more quickly and accessibly than may otherwise have developed unilaterally (Afzalan et al., 2017). In these scenarios, the geoportal acts as a common and general way of accessing the SDI (Hintz & Hantke, 2020) – a sort of Swiss-army knife of informational access. Alternatively, geoportals may also be customized with specialized tools to meet specific end-user needs (Jiang et al., 2020).

See Appendix A.2 for more on commercial geoportals.

2.2 Place and people

2.2.1 Communication

The information age itself is a source of both challenges and potential solutions. Since the turn of the century, all facets of urban life and the structures that support them have transitioned towards the digital and informational. A community as "a system of systems" (Roche & Rajabifard, 2012) has an internal structure (Massey, 1991), with corresponding spheres of influence of its nodes within and outside of these. As quickly

as information and communication technologies (ICT) tools provide new means of characterizing the immediate, physical geographic area of a community node, they also support the digital transmission of ideas and participation to remote parties via direct communication platforms as well as the more public arenas of social media. In short: ”the geography of social relations is changing” (Massey, 1991), with digital connections offering ”unique opportunities to identify and understand information dissemination mechanisms and patterns of activity in both the geographical and social dimensions, allowing us to optimize responses to specific events” (Oliveira & Painho, 2021).

In the course of its operations, a community should facilitate a ”shared understanding of what is happening” within it (Rivera et al., 2020), from planned works to unforeseen incidents. Just as big ideas are evolving through digital channels, so too has the sharing of neighborhood news gone online. Physical proximity is no longer the primary means of passing the latest hearsay. Words are leapfrogging the traditional stoop-to-stoop transmission and sharing information via networking platforms (Evans-Cowley & Hollander, 2010). Following suit, many news channels and government communication departments have incorporated digital distribution strategies, often leveraging social media to engage readers and direct traffic to their channel platforms. This allows not just community eyes on local announcements, but also invites remote viewers to participate (Evans-Cowley & Hollander, 2010).

Stemming from the assumption that ”storytelling is the most effective way to merge meaning and emotions” (World Economic Forum & ScaleUpNation, 2021), a tremendous and increasingly more ubiquitous tool for effective and relatable communication is data visualization (“Giorgia Lupi”, n.d.; Lupi, 2017). Data visualization products and inclusions have migrated beyond the niche tech or business applications to ”a part of the fabric that is modern culture”, threading their way into newspapers, fashion lines and books (Meeks, 2019). Studies indicate that readers prefer pictorial and summary forms of information (as opposed to purely textual) (Evans-Cowley & Hollander, 2010). Visuals can provide additional context, identify changes, and reveal patterns, as well as display and distinguish between relationship types (Shneiderman, 1996), ultimately ”connect[ing] numbers to what they really stand for: knowledge, behaviors, people” (Lupi, 2017). Users, whether they be the general public or decision makers, are expected to have some data visualization literacy (DVL) (Börner et al., 2019). This mutual expectation of information producers, consumers, and actors to present and ingest effective representations is re-enforcing its importance and creating new standards of competencies. ”Every publisher and journalists knows the value of charts and wants more of them” (Google News Initiative, 2018), not only for aesthetic breaks in text blocks and their power to convey complex information memorably, but also the jumps in page views that they generate (“Giorgia Lupi”, n.d.; Meeks, 2019). Simultaneously, academia is establishing DVL frameworks that are being adopted and taught from primary school through higher and continuing education programs.

However, beyond the ability to create compelling visuals to contextualize or communicate important information is the discernment to understand the different insights needed by each stakeholder (Börner et al., 2019). Data visualization can represent answers to the questions like who, what, when, and where by incorporating different kinds of representations (network, topical, temporal, and geospatial analysis, respectively). Maps are increasingly being employed to answer where and related questions as they are more easily interpreted and remembered (Börner et al., 2019). They elucidate spatial relationships using layers of data contextualized by base maps (such as raster images or vector representations) (Baker et al., 2019; Jiang et al., 2020). When presented digitally, maps (much like other charts) provide the opportunity for dynamic exploration and additional insight by visual inspection of elements and their spatial or thematic relations to each other. When evaluating foreign areas, maps provide valuable context by concisely representing proximities and directional situations, versus relying on verbal descriptions that may perhaps be more easily misconstrued (Shneiderman, 1996).

In any case, data should be used and interpreted cautiously. Data records are an abstraction of the real world (Lupi, 2017). Often, visualizations of data omit elements of uncertainty (Meeks, 2019) or are developed prematurely (without proper analysis) or improperly (misleadingly) (Börner et al., 2019; Monmonier, 2018). Further, visualization designers may overestimate the ability of the consumer to interpret them quickly and accurately – one must be careful to display images that can be ingested as intended, without sacrificing nuance of complex issues when it is critical for decision makers (Börner et al., 2019; Lupi, 2017; Zhang et al., 2019).

See additional context about the impact of place on smart communities and public participation in Appendices B.1 and B.2, respectively.

2.2.2 Geospatial intelligence

"There is a big need for spatially referenced data creation, analysis and management" (Bhattacharya & Painho, 2018). In the absence of spatial context, first responders to a natural disaster may be unclear on where or how to apply emergency support (Snyder et al., 2019). Likewise, security teams require spatial data in order to neutralize threats. Primarily associated with defense and disaster relief, geospatial intelligence (GEOINT) uses multi-source data to plan or enact responses to threats in complex environments (Datta, 2018). In additional to earth observation data collected via remote sensing or local environmental sensors, humans and the technologies they wield are passively contributing as potential data mining sources when participating in analyzable social media activity or using applications with geolocated data services. By employing geostatistics and data analysis, these disparate layers can contribute to situational awareness and uncover potential trends (Varanda, 2020). Intentionally and georeferenced data actively shared by civilians or non-governmental organizations also has a

place in GEOINT (Datta, 2018). For example, in the pursuit of international peacekeeping, layers of remote sensing imagery, relationship development, reconnaissance, and open sources may be leveraged to provide a more complete representation of the ground truth. In Portugal, digital news products are monitored and mined for space time patterns, attempting to access overall trends or distinct data that flows directly from those with boots on the ground to the media (Varanda, 2020). However, this can be challenging when considering that written reports are "bounded by restrictions of language and symbols, as well as technical and financial resources" (Baker et al., 2019). Therefore, there appears to be an opportunity for tools to augment the existing textual descriptions of news events that may already include thematic labels with intentional and specific spatial and temporal definitions.

2.2.3 News media

"Everything happens somewhere and some-whence" (Bhattacharya & Painho, 2018), and the reporting on such (current) happenings via large volume distribution channels is what is referred to herein as news media. The way people absorb news has evolved in pace with available technologies. Many users have enjoyed not only a myriad of sources from which to select their desired content, but also a variety of ways to experience it: printed vs. digital written journals, video reports, podcasts, and news aggregators leveraging really simple syndication (RSS) feeds, among others. Especially in digital mediums, embedded links facilitate the discovery of related concepts or deep dives into referred sources, photos illustrate the reported information, graphs allow the evaluation of data products, and personalized accounts filter incoming content to expose the user only to pieces of declared or inferred interest. Even so, opportunities for innovation in the industry abound (Google News Initiative, 2018). Especially within communities, local news (inclusive of official organizations, independent journalism, individual blogs, or social media) supports a repository of local knowledge (Cai & Tian, 2016), contributing to a community intelligence (Xing et al., 2015). The extraction and analysis of the when and where of events can produce valuable information (Bhattacharya & Painho, 2018) without disturbing proximal articles (Teitler et al., 2008). In the context of news reporting, associating geospatial and temporal dimensions to the contents of news stories have several potential benefits (organized below into three categories).

1. Reader experiences: By providing context to readership via the display of information layers on a map, calendar, animation, or other appropriate data visualization, media agencies can provide a better user experience to their readership while keeping them better informed (Teitler et al., 2008). A benefit of digital communication is the ability to reach beyond what we can physically access. However, these same distribution channels may obscure our view into our own neighborhoods (Painho & Pina, 2013). The mapping of news supports the promotion of spatially relevant news

to users, potentially considering both their real-time location and defined areas of interest defined in their profiles (Marshall, 2012). This kind of hyperlocalization tool and "community-oriented digital content" (Leszczynski, 2019) may not only improve reader engagement with the publisher's content (Google News Initiative, 2018; Teitler et al., 2008), but better inform the reader by exposing them to events or issues that are physically close to them (outside of their defined or supposed topics of interest), and may perhaps even promote more engagement with their community (Cai & Tian, 2016; Google News Initiative, 2018).

2. Further analysis of geospatial news stories: By mapping the contents of news, one may identify news deserts (areas receiving little coverage) (Gupta & Nishu, 2020), which can be helpful for publishers to identify new stories to balance their coverage or city managers identifying under-served communities. Other trends, such as topic specific distribution, can be used to gather insights about the interests or challenges facing residents in particular places. "There are many powerful public interest stories out there that will only be discovered if traditional investigative techniques are combined with technology" (Megan Lucero, Director at The Bureau Local (Google News Initiative, 2018)). This information can be fed into various smart city applications, acting as another layer of information that can inform a location-aware response to an event (Roche & Rajabifard, 2012). These events can also be leveraged by GEOINT applications as a layer to help determine "boots-on-the-ground" truth, or identify patterns of deliberate misinformation in near real-time (Imani et al., 2019). It can also serve as the basis for more nuanced research or searching. A spatially enabled platform could allow users to filter news repositories of "textual artifacts of human experience" not only by subject, but also time and place, potentially alleviating the burden associated with reviewing high volumes of documents (Cai & Tian, 2016). These artifacts can also be leveraged by efforts in frame analysis (Hamborg et al., 2019), to inform how these experiences are contrived and communicated.

3. Extracting nuanced place definitions: Once incorporated into a platform, geospatial tools could not only be used as a bonus tool in the distribution of geographical information, but also used as a common baseline and geographical vocabulary from which to collect the "often invisible, descriptive, and vague" knowledge of local residents and further bolster and consolidate community intelligence (Cai & Tian, 2016; Xing et al., 2015).

2.3 Georeferencing news media

2.3.1 Geoparsing unstructured documents

As of 2019, approximately 60% of the data was associated with a geolocation (Karimzadeh et al., 2019). Though technology permits this percentage to grow as more data may automatically be referenced upon creation, many documents still do not have associated

georeferences, neither for the document itself (a publish location), for the contents as a whole (focus location) or the sub-contents (place mentions within a document that may vary across events) included in that piece of data (Gritta et al., 2018; Halterman, 2019). For these, extracting the "unambiguous representation (or footprints)" (Cai & Tian, 2016) of locations is not a straightforward task, neither for humans nor automated processes (Lee et al., 2019). In best case scenarios, this geo-association involves the geocoding of structured references that have discrete associated geocoordinates (Gupta & Nishu, 2020; Hamborg et al., 2019). More likely, however, the process requires the geoparsing (automated geotagging of unstructured textual descriptions) via toponym recognition and toponym resolution of potentially ambiguous or relative place mentions (Halterman, 2019; Karimzadeh & MacEachren, 2019). Then, the location of the contents may be extracted (Imani et al., 2019) (geocoding), and perhaps, depending on context, the appropriate geographic focus can be determined (Silva et al., 2006; Teitler et al., 2008). Recognizing toponyms requires the identification of and distinction between names of entities, most often achieved via the process known as named entity recognition (NER) (Gritta et al., 2018; Silva et al., 2006), which is subsequently refined to only the location candidates (Imani et al., 2019).

A gazetteer is a geographical index relating toponyms to descriptors, and usually to location via geographical coordinates. One can compare phrases (potential toponyms) in the text against a gazetteer to identify potential matches (Lieberman et al., 2010; Silva et al., 2006). Manual geo-annotation services allow users to georeference digital artifacts by searching place names, points of interest, or street addresses in a textual format from online gazetteers, and/or permitting the definition of a feature directly to a map interface (Elwood et al., 2012).

However, accurate resolution of geoparsing may not be straightforward. More elaborated gazetteers contain more ambiguity as their coverage and detail grows (Karimzadeh et al., 2019). There is also a challenge of resolving multiple names (including various endonyms and exonyms), which has prompted various concerted efforts from specialists in related areas to standardize place names, such as The United Nations Group of Experts on Geographical Names (UNGEGN) which compiles nationally composed gazetteers (Witschas, 2004). Even so, the vernacular terms used in non-adherent documents still require resolution, which is sometimes resolved by the association of known alternate names (see additional information on toponym disambiguation in Appendix C.1). Granularity is another challenge. Many gazetteers associate place location to a point, whereas areas may be more appropriate to smaller scales. Consider the city of Lisbon, Portugal: if attempting to uncover patterns within the city, resolving all Lisbon-related areas to the coordinates (38.7223° N, 9.1393° W) will not reveal any great insights about details within the city (Cai & Tian, 2016). Increasingly, VGI is being leveraged in local-level applications to take advantage of the colloquial terms and commonly understood (yet not officially delineated) places. While this tackles the uniqueness of localities, this too may suffer from incompleteness or inconsistent

coverage (Cai & Tian, 2016).

One relevant and well-cited gazetteer example is OpenStreetMap, which leverages VGI to compile detailed and updated spatial data throughout the world. The resulting data products are open and licensed under the Open Database License (ODbL) by the Open Street Map Foundation (OSMF). A previous study leveraging its data as a gazetteer reported that it includes more granular and accurate places than other gazetteer options and recommended its use in future works (Al-Olimat et al., 2018). Another important example is GeoNames.net, which has been selected in previous geoparsing works because of its data sources, unique identifiers, "extensive coverage, quality, inclusion of metadata (such as alternate names and geographic hierarchical information), and frequent updates", considering it "the richest gazetteer available at the time" (Karimzadeh et al., 2019). Between the two, it appears that OpenStreetMap is more appropriate for local applications, while GeoNames is better suited for international studies.

In toponym resolution, candidates are compared to all location name entries in a gazetteer. This, however, is rarely sufficient for accurately distilling the appropriate location. Vernacular text may refer to locations in alternate ways, including shortening. Context also plays an important part in accurate geoparsing. In an attempt to disambiguate which potential match in a gazetteer is intended by a spatial toponym, geocoders may employ contextualization tools building from other queues in the text. For example: if multiple places are being referenced, common threads (hierarchical level, clustering, regional adherence) may be leveraged to identify probabilistic matches (Al-Olimat et al., 2018; Gritta et al., 2018; James, 2020; Lieberman et al., 2010). See Appendix C.2 for existing geoparsing tools.

2.3.2 Event extraction

Beyond the geoparsing of texts is event extraction, which attempt to recognize not only the place but also actors, time, and details related the event (Halterman, 2019). By extracting and then mapping news events, articles and their contents can be summarized, clustered, and aggregated (Hamborg et al., 2019). In areas of conflict or political variability, understanding exactly when and where events have occurred can have GEOINT related benefits (Lee et al., 2019), ideally in near real-time. It can also improve accessibility to consilience, which may contribute to further local engagement and development (Cai & Tian, 2016; Elwood et al., 2012). In any case, this event extraction is foundational for further processing and knowledge extraction (Hamborg et al., 2019). Precisely extracting location is critical for its effectiveness (Gupta & Nishu, 2020).

Journalist reporting often employs phrases that textually describe an event ("explicit event descriptors") to answer the who, what, when, where, how, why. These are

intended to be understood by a human reader, but are also available to artificial ingestion methods (Hamborg et al., 2019). To determine the where, the process of toponym recognition and toponym resolution is followed by the determination of each location name as either the focus or a periphery location (Lee et al., 2019), both of which are often different than where the document describing it originated (Snyder et al., 2019). Moreover, to be useful at local scales, events often require more specific location information than existing administrative boundaries (Halterman, 2019). In fact, due to the nature of human events (especially in the case of conflict), locations may spread to areas uncontained by any particular definition in a gazetteer entry (Lee et al., 2019). Similarly, in addressing the when, articles may not definitively spell out the time frame associated with an event (Hamborg et al., 2019). Instead, temporal references may use the date of publish as an anchor point, employing relative terms such as "yesterday". They could also require additional knowledge of peripheral events, such as "during the last week of school", or reference other named entities and function words. Likewise, just as areas can be fuzzily defined, so too may time frames be blurry. Start or end dates and times of events may be nebulous, without an obvious way to translate an event start or end time to a calendar or clock.

News is generally written with a specific audience in mind, and authors will use words consistent with their shared vernacular of their readers (Lieberman et al., 2010), and may reference places in relation to a locally recognized point of interest (POI) (Lee et al., 2019). As opposed to general gazetteers intended to suite all possible needs, local lexicons (collections of geographically proximate toponyms) focus on particular sources and audiences in order to create better performing toponym resolution (Lieberman et al., 2010). In ongoing stories, more recent, potentially under-specified toponyms can be assumed to reference previously used and well-specified toponyms, as part of an ongoing conversation via published documents. The gazetteers for local news, then, should be hyperlocalized to their content (Cai & Tian, 2016). However, it should be noted that spatial proximity does not mean linguistic proximity. Different news agencies often exhibit different writing tendencies in terms of content, vocabulary, style, and emotions (Imani et al., 2019).

2.3.3 Automatic geo-annotation

Consider the tremendous number of existing digital news articles (not to mention the articles published prior to the transition to digital platforms in the 1990s and early 2000s or those that continue to only be available in print form). The repositories of textual event data is vast and growing constantly with every new article published. The task of geocoding existing corpora of unstructured text documents (news archives, social media posts, websites, etc.) will likely require highly detailed and well-trained

machine learning programs that can augment existing data with derived location associations, as human geo-annotation resources are slow, expensive, limited, and inconsistent (Halterman, 2019; Lee et al., 2019). Especially in the case of imprecise geographic definition, relative geography phrases can be used with anchor toponyms to extrude the true locality of an event (Lieberman et al., 2010). By processing nonstructured documents in the way they were intended to be ingested by human readers (linearly), resulting heuristics and natural language processing (NLP) methods may lead to more precise results (Lieberman et al., 2010). See Appendix D.1 for examples.

In addition to the well-described challenges of accurate toponym identification and disambiguation, a common theme among the literature describing existing tools is insufficient training data against which to test the developed methods (Imani et al., 2019; Karimzadeh et al., 2019). Likewise, many of these methods do not translate well to other topics, scales, locations, or document types without considerable retuning, let alone retraining. Many of these also only resolve to point coordinates, instead of areas which may better represent the locations of events of interest, relevant in both visual inspection or during further aggregation or processing to extract more information representing ground truth. Almost all of the options are only viable in a single language, usually English. Another challenge is that evaluation metrics differ from study to study: false negatives and/or false positives may not be considered in precision metrics, which may be due to incomplete gazetteers (Lieberman et al., 2010). Most importantly, all of these attempt to extract event location with as much precision as possible and minimizing errors. Of course, an ideal situation is to have exact locations already associated.

2.3.4 Manual geo-annotation

Though the task of retroactively associating locations to text documents is desirable for many applications, it is also, as was previously described, prone to ambiguities and mis-association. To minimize these, human annotators are required to create subsets of data for testing and training of the geoparsing models. This kind of ground truth is critical for the improvement and evaluation of such automated processes. See Appendix D.2 for examples.

Manual solutions offer a potential for higher accuracy in geoparsing news articles, however they introduce the cost (monetary and temporal) of introducing humans into the process. Also, though ground truth is considered how a human annotator would parse a document, if the annotators are not the author or subject experts, their association may be inconsistent, imprecise, or inaccurate (Karimzadeh & MacEachren, 2019; Lee et al., 2019). This is augmented if human annotators have their own context and therefore different understandings of place, resulting in different geolocations assigned from annotator to annotator (Halterman, 2019).

2.3.5 Author geo-annotation

Therefore, it is also interesting to explore the possibility of intentional geo-annotation of a documents' contents by the author of the document him or herself. Though this manual geo-annotation will almost certainly require addition tools, time, and effort on the part of the author, it will provide the exact ground truth to which the contents can be better understood and automated efforts are compared against. A resource that provides a way for authors to define custom places, i.e., define or design their own geocodes for a new gazetteer entry, may have specific and broader impacts. For the author's application, it may provide the flexibility for the author to define exact areas at a higher granularity than provisioned in existing gazetteers. This kind of progressive geocoding may also incrementally enrich local gazetteers as a new source of VGI (Cai & Tian, 2016), and promote search engine indexing for local results (Lieberman et al., 2010; Marshall, 2012). Existing annotation tools may not reflect the vernacular means of describing place in any particular locale (Cai & Tian, 2016). There is also the challenge of the mutable nature of digital content: any article may change or disappear after publishing, which is challenging to account for if its event attributes have been established by a third party (Teitler et al., 2008). By enabling a document's author to directly associate geolocation(s), many of the challenges associated with 3rd party annotation may be avoided. It should be noted, of course, that as most news agencies and their staff do not have formal geographic training, the resulting features should be considered within the scope and with the considerations of VGI.

Appendix D.3 reviews some existing examples. From these, its clear that existing tools to support journalist geo-annotation of articles are lacking. Journalists have very few options for geo-associating their own content, and these are all limited to one or multiple point locations. However, the sheer number and valuable results of efforts dedicated to the extraction of exact location of the news demonstrates a very clear need for true (which often means two dimensional) association of location, from which further analysis and GEOINT applications can draw.

3

Methodology

3.1 Justification

3.1.1 Concept

The following project (nicknamed Apregoar, a Portuguese word meaning 'to proclaim') is a foundational, proof of concept web app that seeks to demonstrate the possibilities of intentionally associated TA and SAs to news stories by media agencies for an improved user experience for traditional readership, as well as improved searching capabilities for researchers and informational dashboards for monitors. Through various tools utilizing the underlying geospatial database, a variety of users may interact with news media in new ways to glean additional insights about relevant places, histories, or previously obscured geospatial patterns. The myriad of potential functionalities, though intimately connected by a communal database and shared system architecture, provide such different user experiences that they can be considered different informational products. This proof of concept, therefore, focuses on only a subset of those possible to demonstrate the core hypothesis: that the association TAs and SAs to news stories will provide additional and valuable information. At the conclusion of this work, the prototype is available for testing by variety of users, drawing constructive criticism, and ultimately serving as a base from which future, more fully featured platforms may evolve.

Of the key drivers of geoportal advances (including scientific geospatial projects and applications and international organization) commercial and governmental drivers are most applicable to this project, as they have the greatest voice and interest in conveying more clearly the spatiality of happenings within a hyperlocal community (Jiang et al., 2020).

3.1.2 Key functionality

The following functionalities stem from the same geodatabase, though each can be implemented independently. Each of these informational products attempts to illustrate

a potential dimension of a such an online geospatial tool (derived from the user story of Appendix E.2).

1. A spatial database of incidents that supports the association of SAs, TAs, and ThAs.
2. A geo-annotation tool that allows journalists to place and time the described happenings (instances) of their articles.
3. A context map allowing publication readership to visualize, per article, the location of its subject matter.
4. An exploration tool allowing readers or researchers to filter by SAs (one or multiple defined places or via drawn definition of the study area), TAs, and or ThAs.

3.1.3 Requirements

The Apregoar system architecture reflects the needs of the browsers, researchers, monitors, and agencies anticipated to use the various functions of the tool, as per the earlier description and the following specifications:

- The system should support all create, read, update, and delete (CRUD) operations, though most tools will only require reading of selected database records. This should include text, numerical, array, datetime, and spatial types.
- The system should be implemented with no- or low-cost maintenance.
- The system should support open source applications, and therefore utilize openly licensed tools as much as possible.
- The system targets non-professional GIS users, and therefore should be straightforward to use, leveraging familiar GUIs as much as possible.
- The system will use spatial, thematic, and temporal filtering, and will therefore leverage data frames that can support this type of rapid processing. Likewise, definition of such filters in an easily understood format are necessary.
- Users should be able to define polygons. This is applicable both in the georeferencing of incidents (authors may define one or multiple polygons defining the location of an incident), as well as for defining spatial search areas for use in filtering georeferenced articles.
- The main search results will be presented in map and list formats. Therefore, the system must support this type of data visualization.

These requirements are derived from the key functionality of Section 3.1.2, user interviews, and best practices derived during the literature review (Section 2.3).

3.1.4 Distinction from previous work

Unlike many of the example provided in the literature, this effort seeks to georeference articles instead of social media posts, most notably "tweets". Tweets or other specific data types include structured organizations and data APIs from which automated programs attempt to derive sentiment and/or relation to particular events (Snyder et al., 2019).

The movement to incorporate citizens as sensors is important and powerful, but we have jumped over public and commercial sources of information that can not only contribute to but contextualize citizen feeling towards a place. Public and private data sources of events are being underutilized – the content exists but needs to be georeferenced in order to be better accommodated by citizens, public management, or private enterprises. Citizen knowledge of a place is drawn from both anecdotal experience as well as learned (read) information from third person sources (newspapers, reports, etc.). The association of place, especially two-dimensional definitions, to news sources can be studied for its influence on public opinion and determine how the spread of news affects public opinion.

This project also fills a different niche than the projects attempting to parse a variety of information (SAs, TA, and ThA such as sentiment, volatility, etc.) from international journals and articles. While immensely valuable for a host of applications, the inherent nature of automation and post-processing (of articles or tweets) makes this method prone to inaccurate results. By incorporating not just human-in-the-loop but author defined association, this project seeks to develop a novel standard for allowing a journalist to explicitly assign TAs and SAs to data records which can then be used as highly accurate input for future extrapolation of causality or trends within a community. The resulting data set should, therefore, minimize mis-association of time and place, and create a new standard for ground truth geotemporal attribute definition.

Further, most automation projects at the international level attempt to associate location to a city level. While this level of granularity is likely sufficient for most international applications that seek to evaluate trends on a grand scale, it brings no further insight to hyperlocal exploration. Local officials interested in monitoring on a parish or even neighborhood (and therefore not administratively defined) level will gain no further insight from associations to the city as a whole. Likewise, as is clear from the previous discussion on place and spatiality, users bring a variety of realities in association with places. Among the different types of users of the city (at the work vs. play vs. live levels), the name of place will be understood differently. Moreover, neighbors within a particular parish may define the same neighborhood in different ways, or use different language to describe it. The Apregoar toolset should allow journalists with a clear understanding of where an event (as the subject of his or her article) is occurring (or was or will occur) to define its boundaries outside of common understandings or administrative definitions. This context provides a common definition of place that can

be visually understood and persist beyond changing borders or evolving names.

Moreover, though much previous work has focused on point definitions of place, nothing happens in a location of zero dimensions (a point). Therefore, the Apregoar toolset currently only supports polygon definition. In cases in which points tend to be more appropriate, users are invited to draw sufficiently small polygons, representing a single building or even a sub area of this. By inviting the design of custom areas (which can use existing boundaries as templates from which to draw) it also pushes journalists to think beyond existing areas and to carefully consider whether indeed their spatial description applies to entire administrative boundary, or perhaps is less completely and binarily affected across that space. It shoves off some of the rigid definitions already assigned to areas that fall within physically or politically defined areas, and permits new identities to formulate across or within these predefined areas (Baker et al., 2019). For analysis or visualizations requiring point data, this transformation can be processed on the fly (Brown & Pullar, 2012).

Just as a reader's lexicon develops over time, so should the associated gazetteer. Manual specification allows the journalist the opportunity to name and define these flowing areas as their and general understanding of an event's location changes over time. Say, for example, that a military base expands or moves – the same named thing within a city could come to mean different footprints at different times. The flexibility to adjust on the fly is critical to the success of this type of initiative (Lieberman et al., 2010). These custom, progressive gazetteers, beyond contextualizing the news and providing a spatial database of associated place, can also be used in and of themselves to explore the patterns in understanding of how a place is named, external to any event that has occurred there. For example, an area may be frequently referred to colloquially with one name, though technically it may be associated to another place. This element provides an indirect opportunity to study place within the study area, external to any official news communication.

3.2 Study area: Lisbon, Portugal

Lisbon is the capital city of Portugal, located in the central west area of the continental country, just north of the River Tejo. Just over half a million people reside within the municipality of Lisbon, a total of 2.9 million in the Área Metropolitana de Lisboa (AML) (a NUTS II region of 17 municipalities both north and south of the River Tejo (Eurostat, n.d.)), and 2.3 million in the overlapping Distrito de Lisboa (DdL) (16 municipalities, entirely north of the river) (Institudo Nacional de Estatística, 2021). The union of these two administrative boundaries (Figure 3.2) form the area of study for this work; Apregoar study area (ASA).

Table 3.2 shows the breakdown of various definitions of Lisbon. In 2012, a nation wide effort redefined the freguesia boundaries. This consolidated many existing areas

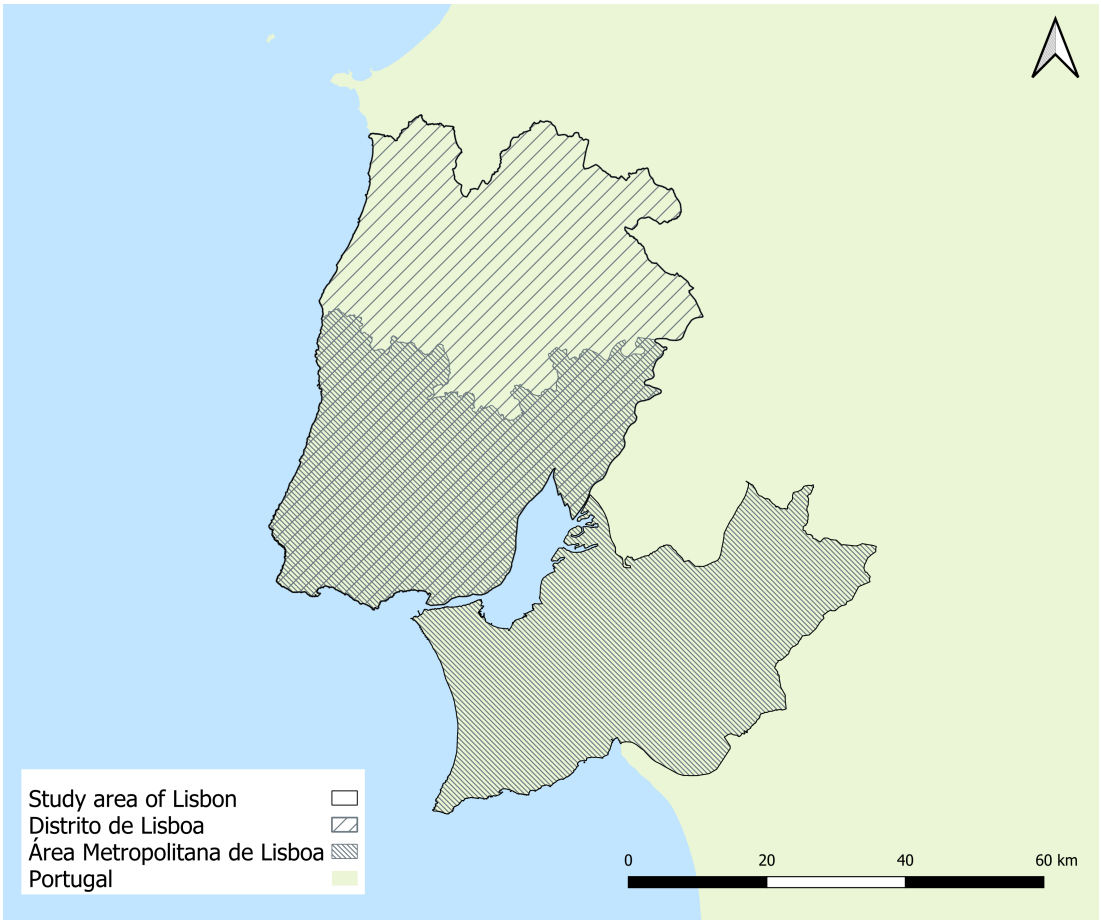

Figure 3.1: Study area of Lisbon, the union of the Distrito de Lisboa and the Área Metropolitana de Lisboa

and redefined boundaries of others, greatly altering the official nomenclature and administrative frontiers. Of course, the previous names and definitions may live on in the cultural memory, contributing to reference ambiguity.

Definition	Districts	Municipalities	Freg 2022	Freg 2011
Continental Portugal	18	278	2882	4050
ASA	2	25	171	284
DdL	1	16	134	226
AML	2	18	118	211
CML	1	1	24	53

Table 3.1: Breakdown of study area

Within the Lisbon area, a variety of organizations serve communication media to the population, including such commercial endeavors as Público, Diário de Notícias, Jornal de Notícias, Observador, and at least ten other major news sources. Many of these have

at least partially transitioned to an online presence, with some available exclusively online. A Mensagem, for example, is a recent addition (as of February 2021) to the news scene in Lisbon, is available exclusively online with some stories in the audio podcast format. This particular source is specifically aimed at hyperlocal information for the Lisbon community. Additionally, the CML provides regular bulletin and news, as do many of the parishes within the municipality and beyond. News is heavily embedded in the culture, with many cafes and restaurants often streaming news channels and printed media available at kiosks at corners throughout the area. There is also a history of pirate radio in Portugal which contributed to the decentralization of speech (Bonixe, 2019). Portugal more generally also has taken an interest in innovative journalism; João Palmeiro, president of the Associação Portuguesa de Imprensa (Portuguese Publishers Association), chairs Google's Digital News Innovation Fund (DNI Fund), for quality journalism in Europe, which both supports Portuguese media projects and includes commercial and academic Portuguese partners (Google News Initiative, 2018).

The hilly and water-lined layout of the city lends rich and distinct character to neighborhoods throughout the metropolitan area. Some of these cultures have persisted through decades, like the areas of Alfama and Alvalade, while others are constantly forging new identities and even rebranding themselves to invite new stories to be told, such as the Martim Moniz and Marvila areas. This, then, makes Lisbon an interesting case study to test a spatial news visualization, to see if the heterogenous personalities will be reflected in any geospatial patterns.

More interesting than the freguesia level is the opportunity to associate stories by neighborhood, which are more granular than parishes and also less reliant on administrative boundaries, perhaps better representing the lived reality of city users. For example: within the same parish, demographics may differ wildly. Politicians may use different clothing, language, and tones when interacting with the people from different neighborhoods, demonstrating variations in spatial autocorrelation on a very fine level. This, of course, presents a challenge as there is no formal definition of neighborhoods as they are neither officially recognized nor static. Some applications have defined approximations of these (see the differentiation used by Idealista in their spatial searching functionality, Figure 3.2), but these are not publicly available. Other opportunities to approximately define these areas include using existing gazetteers (such as Open Street Map) which may include textual association with neighborhoods in their descriptions, and create minimum bounding areas around these to indicate general areas.

Since there is a lack of definition in this area, perhaps there is an opportunity for the progressive gazetteer developed in this application to inform future approximations of neighborhoods. If, for example, enough stories define instance areas as neighborhoods, over time the overlap of these could be used to indicate fuzzy neighborhood definitions for this or other purposes.

Figure 3.2: Vector approximation of neighborhoods in Campolide (Idealista, n.d.)

3.3 Data collection and preprocessing

3.3.1 News corpora

Of the available news sources, the following were selected as a representation of sources of event reports for the months of January, February, and March of 2022. The resulting corpora incorporate several public organizations and a private news media source to inform the organization rubric and data model. As historical values, these are used to develop and test the application.

CML is the governing body of Lisbon municipality, supporting its development and initiatives. The council oversees areas related to the environment, social rights, culture, innovation, mobility, etc. at the city level. The municipality is currently broken into 26 subdivisions of freguesias. It communicates with the population via:

- active social media campaigns (including via Twitter, Facebook, Instagram, LinkedIn, and Youtube),
- periodic publications on a variety of themes,
- a regular municipal bulletin communicating its resolutions,
- downloadable announcements, notices, and warnings
- online news

This work uses the ”notícias” (news) section of CML website in the geocorpora, as its blog format is most conducive to planned iterations of the tool (Câmara Municipal de Lisboa, n.d.).

Junta de Freguesia de Campolide (JFC) (14.8K residents (Institudo Nacional de Estatística, 2021), 277 hectors (DGT, 2022)) was selected for inclusion in the geocorpora as the freguesia in which the University of NOVA IMS is located and has previously collaborated with the innovation department. Said department has also expressed interest in such a tool for better management of the territory and communication with their constituents. In the most recent election (September 2021), the leadership of JFC changed and the latest regime is currently undergoing a rework of its communication materials, including its website. Therefore, the previously available news blog is no longer available or being updated for the time being. In its stead, the geocorpora currently includes a subset (due to time constraints) of posts published by the JFC Facebook page (Junta de Freguesia de Campolide, 2022) during January through March 2022. This also provides an opportunity to explore how such a tool might eventually interact with existing social media tools.

Junta de Freguesia de Campo de Ourique (JFCdO) (22.1K residents (Institudo Nacional de Estatística, 2021), 165 hectors (DGT, 2022)) was selected for inclusion in the geocorpora as the freguesia in which the author resides and therefore has on-the-ground understanding of the territory in which to associate instance location. This Junta has released news via their website through the study period, however only in a very small amount (1-2 per month) (Junta de Freguesia de Campo de Ourique, 2022).

Junta de Freguesia de Esrela (JFE) (20.3K residents (Institudo Nacional de Estatística, 2021), 560 hectors (DGT, 2022)) was considered for inclusion in the corpora as the freguesia in which the author previously resided and therefore has on-the-ground understanding of the territory in which to associate instance locations. This Junta has released news (at least weekly, (Freguesia de Estrela, 2022)) and events (intermittently). Though not incorporated in the geocorpora (due to time constraints), the format of their news and available attributes are considered in the formulation of the Apregoar data model and resulting web app tools.

A Mensagem (AM) is a digital newspaper "about Lisbon, of Lisbon, and for Lisbon" (A Mensagem, 2022). It is particularly concerned about updating digital news practices to better serve the local community, highlighting relevant social initiatives and issues, as well as encouraging community participation. At the onset of the development of this thesis, Catarina Carvalho, founder and editor of a Mensagem shared her insights on the publication process of digital news publications, as well as interest in developing such a tool that would help their readership spatially contextualize their stories. Ideally, future development would involve closer collaboration to implement a larger scale test, incorporating interaction with the geodatabase into the existing publication process of this journal via a WordPress plugin to the existing process. A Mensagem publishes multiple stories daily, usually with one or more instance localizations in the Apregoar study area, applicable to the entire country, or (very rarely) beyond.

Público is a digital and print newspaper serving Portugal, representative of more traditional and established publications that incorporate news at local to a worldwide

scale. Público provided the results of a query of its articles that might contain Lisbon or local news for the month of October 2020 (Appendix E.1). The results were analyzed for the availability of relevant ThA, TA, and SAs. The results of this analysis were considered in the formulation of the Apregoar data model and resulting web app tools.

The above sources form the informative corpora, sources of news stories (each associated with a single article or post uniform resource locator (URL)) from which relevant insights and attributes were extracted to inform the data model and ultimately the various tools of the web app. Those articles entered into the Apregoar geodatabase are considered the geocorpora, which serve as the test data for the Apregoar system.

Story attribute	type	CML	JFC	JFCdO	JFE	AM	Público
Title	Text	Yes	Maybe	Yes	Yes	Yes	Yes
Summary	Text	No	No	Maybe	No	Maybe	Yes
Publish date	Date	Yes	Yes	Yes	Yes	Yes	Yes
Section(s)	Text	Yes	No	Yes	No	Yes	Yes
Tags	Text	Yes	No	No	No	No	Yes
Author(s)	Text	No	No	No	No	Yes	Yes
Link	Text	Yes	Yes	Yes	Yes	Yes	Yes
Instance(s) attribute		**Type**	**All informative corpora sources**				
Place name		Text	Maybe				
Place definition		Geometry	Maybe				
Place description		Text	Maybe				
Start date		Date	Maybe				
End date		Date	Maybe				
Start time		Datetime	Maybe				
End time		Datetime	Maybe				
Temporal description		Text	Maybe				

Table 3.2: Considered attributes of informative corpora

All story attributes included in Table 3.3.1 are copied directly from the source materials if available. In some cases, the ”section” is inferred from the publication’s website when displayed (such as ”Bairro” on the A Mensagem page). In the absence of tags, sometimes the themes are extrapolated and defined by the researcher to enhance the search capabilities of the system.

Instances refer to each geotemporal definition(s) of a story; that is to say: each distinguishable pair of ”where(s)” and ”when” associated with an event. Each story may have multiple instances associated with it, and each instance may include one or more area. Instances may not be associated with more than one story, but geolocations (gazetteer entries, predefined or custom) may be associated to the instances of multiple

stories. Instance attributes (TAs and SAs) may or may not be explicitly stated. In the case of administrative boundaries, these can be mapped to existing gazetteers. Otherwise, in lieu of access to subject experts (author of the article, subjects of the article, etc.), the researcher has attempted to extract the location from text descriptions within the article body, searched via google maps and street view, and/or scoured the web for indications or definitions of place (as in the case of neighborhoods that have no official boundary yet are within the cultural understanding). Time attributes are likewise extracted from the text body. When necessary, relative descriptions (such as "yesterday", "on his birthday", etc.) are translated to dates and, if appropriate, times.

	CML	JFC	JFCdO	AM
`Stories (278)`	92	39	4	143
`Instances (433)`	123	45	5	260

Table 3.3: Geocorpora

3.3.2 Basemap

Basemaps are included to orient the journalist or reader, improve the browsing user experience, as well as indicate references for defining geometry, either in the case of establishing the location of an incident or establishing a search area relative which to filter instances. The basemaps used in the Apregoar web app are produced by Open Street Map and Stamen (OpenStreetMap contributors, 2022; Stamen Design, n.d.), both of which include additional base features (such as roads, parks, blocks, etc.) to provide additional context to the user.

No preprocessing is required for the basemaps.

3.3.3 Gazetteers

Apregoar uses gazetteers for several functions:

1. Direct assignment of instance locations, Example: CML announcing changes to the garbage collection schedules within Lisbon, associating the instance to the administrative boundary of the municipality of Lisbon.

2. As a reference when creating custom locations to which to associate a news instance. Example: because definitions of neighborhoods are not available, an author describing the Boavista neighborhood may search gazetteers matching the name to serve as reference points or areas from which to define the localization of the news instance.

3. Spatial filtering of news instances while interacting with published articles to support disambiguation and relevant results. Example: limiting the search results

to only include stories with instances that intersect freguesia de Santo António so that a search for 'Pombal' doesn't include instances occurring in the freguesia of Pombal in Leiria.

4. Automated spatial enrichment of custom defined instance locations to improve filtering by administrative boundaries. Example: if a journalist draws a custom polygon representing the spatial aspect of a news instance occurring in Praça Marques de Pombal, its geometry is compared against the existing gazetteer entries and, in the case of intersection, the instance is associated to the intersecting gazetteers, including freguesia Santo António, freguesia Avenidas Novas, freguesia Coração de Jesus (a retired freguesia), Lisbon municipality, DdL, AML, etc. When a reader or researcher defines administrative boundaries of interest during a search, this pre-processing supports more efficient results (versus performing spatial queries on the entire geocorpora at time of search).

5. A repository of custom, journalist defined places that is made available for re-association to new instances. Example: A journalist is chronicling a particular story in several pieces that refer to a block in the neighborhood of Campo de Ouriqe that the journalist had previously defined. He (or anyone else) may use this custom definition without needing to redraw the boundary.

6. A repository of custom, journalist defined places that may be used to extract new understandings of place, heretofore obscured. Example: The ugazetteer (short for user gazetteer, or custom gazetteer) could be used to identify hotspots of attribute values or textual descriptions and provide new insights into the cultural understanding of the Lisbon area.

The Apregoar project uses several existing geospatial resources to feed its gazetteers. Pre-existing gazetteer data include those from official (available through government open data portals) sources, such as administrative boundaries (Table 3.2, current and retired), as well certain environmental definitions. The latter (specifically parks and green areas of Lisbon) were included in this prototype due to frequency of which they were referenced by the geocorpora.

Source	`Name: Contents`	Geometry	SRS
Hosted in Apregoar spatial database			
DGTerritório	`CAOP2017: Administrative boundaries`	MultiPolygon	EPSG:3763
DGTerritório	`CAOP2011: Administrative boundaries`	MultiPolygon	EPSG:3763

DGTerritório	CAOPv1: Administrative boundaries	MultiPolygon	EPSG:3763
CML Geodados	Espaços Verdes: Green areas in Lisbon	MultiPolygon	EPSG:4326
CML Geodados	Grandes parques e Jardins de Lisboa: Parks and gardens of Lisbon	MultiPolygon	EPSG:4326
	Connected via API		
GeoNames	A geographical database of place names	Point	EPSG:4326
Nominatim	A search engine for OSM data	Point, Polygon	EPSG:4326

Table 3.4: Existing gazetteers

To prepare the hosted gazetteers, each data source is manipulated using QGIS3 on a Windows 10 operating system. Each data layer is loaded (either via download from or connection to the appropriate geoportal) into the program and, if necessary, transformed to EPSG:4326 (as the SRID of GPS and the default of many geospatial manipulation tools, this is expected to provide the most straightforward user experience and future project expansion). The various Carta Administrativa Oficial de Portugal (CAOP) layers and GeoNames include records throughout the entire continental region of Portugal, and therefore need to be reduced to only the spatially relevant features and attributes.

The CAOP layers were filtered to the relevant municipalities (those constituting the study area). Additional features were creating by leveraging the DICOFRE, a 6 digit code indicating the district (the first two digits), conselho (middle two digits), and freguesia (final two digits) of each entry. To ensure that each freguesia was associated with a single record, a dissolve applying to the DICOFRE attribute was applied to each dataset. Then, each dataset was dissolved to the first four digits (indicating district and conselho) of the DICOFRE (now DICO) to establish a boundary for each municipality of each CAOP dataset. This process was repeated for the DdL by extracting and dissolving to the DI (first two digits of the DICOFRE), where DI = 11 (the code associated with DdL, and again where the DICO values match those associated to the municipalities constituting AML. Finally, the entire dataset was dissolved into a single feature indicating the boundaries of continental Portugal. These features were saved with attributes describing the original identifier (ID) (if exists, such as DICOFRE, DICO, or DI) and source of the dataset, a descriptive type (ex: "freguesia"), name, and its geometry.

No spatial modification was required for the CML Geodados datasets, however they were associated with the same attributes as the CAOP layers (original id, source, type, name, and geometry).

The administrative features and CML managed points of interest were loaded into the PostgreSQL egazetteer table via QGIS. In fact, any data preparation or local testing utilized QGIS, selected for its adherence to open standards, prior experience, and support in literature (Sami, 2019).

3.3.3.1 Connections to existing databases

Besides loading and transforming existing features into a locally hosted gazetteer, existing gazetteers may also be accessed via web APIs that permit the extraction of remotely hosted data via HyperText Transfer Protocol (HTTP) over the internet. This kind of remote connection permits the inclusion of dynamic gazetteers, especially VGIs that may include more colloquial or functionally relevant entries that can be associated with news instances. This work leverages the GeoNames and OSM Nominatim gazetteers, both of which include point geometries which can be used for contextual referencing, which has been noted as important for assisting in the disambiguation of human annotators when selecting toponyms (Karimzadeh & MacEachren, 2019). The Nominatim gazetteer also includes polygon features to which the news instances can be directly geotagged. In their respective API calls, the results are limited to the envelope of the study area, defined by its maximum and minimum latitude and longitude coordinates: North (y_{max}) = 39.83801908704823, South (y_{min}) = 38.40907442337447, East (x_{max}) = -7.74577887999189, and West (x_{min})= -9.517104891617194. As these data connections are initiated by user interactions in the process of using the tools, all manipulation of the results is done by the web app itself. Results are requested in Javascript object notation (JSON) format to facilitate this manipulation in both the front- and back-ends (JavaScript (JS) and Python, respectively).

Additional sources, including further information about transportation stations (metro, train, bus, etc.) or other managed areas may be included in the future to make searching more robust and easier to use.

3.4 Development

3.4.1 System architecture

The Apregoar system architecture was developed to support the functionality and requirements established in Section 3.1.1. It uses a multi-tier architecture that permits the logical and physical separation of the functionalities of the web app. This type of organization has been long utilized and is still a common practice in the field (AWS, 2021; IBM Cloud Education, 2020; Panchaud & Hurni, 2018).

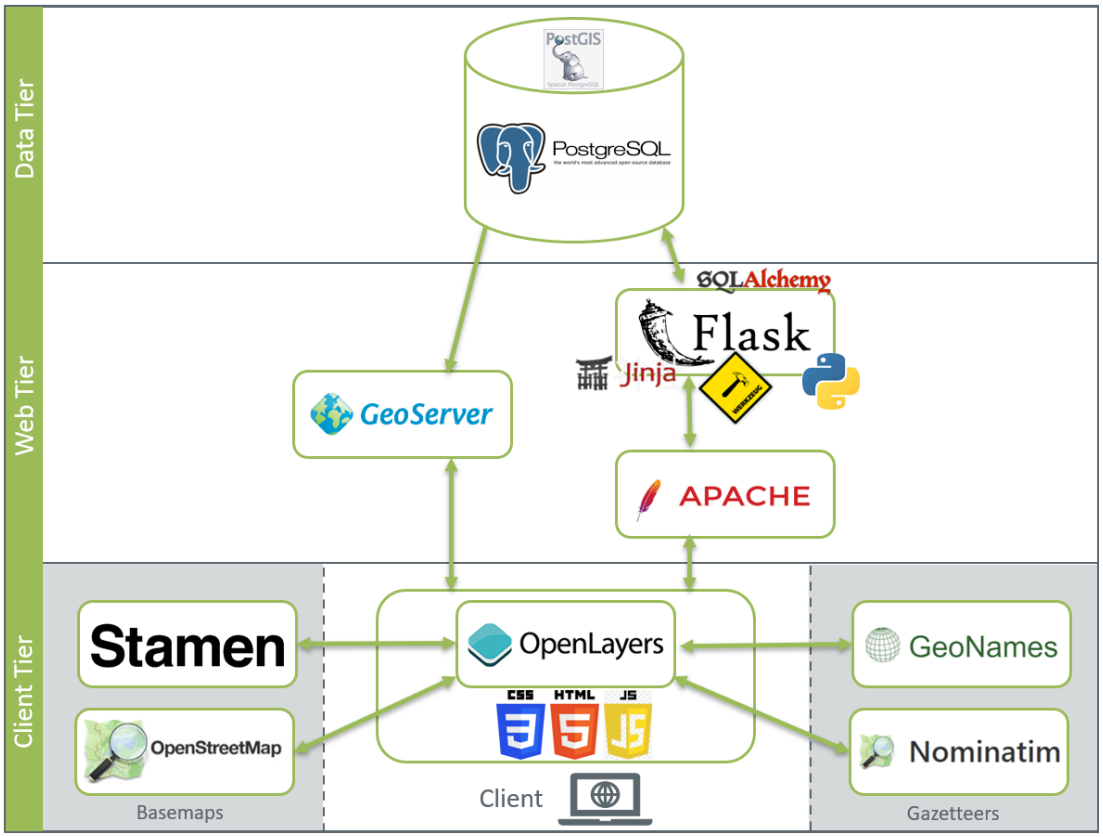

Figure 3.3: Apregoar system architecture

3.4.1.1 Data tier

PostgreSQL and PostGIS: An event model requires spatio-temporal and attribute elements, with the former requiring specific formats and processing for data manipulation and storage. For this reason, the data storage leverages PostgreSQL (version 13.2 on a Windows 10 operating system) with a PostGIS (version 3) extension for its vector data storage capabilities, support in literature (Bhattacharya & Painho, 2018; Oliveira & Painho, 2021; Sami, 2019; Teitler et al., 2008), support for remote connection via Python using related packages, and the researcher's previous experience with the platform.

3.4.1.2 Web tier

Flask: The back-end is based on the Flask web development platform, chosen for its lightweight implementation and Python programming language. It has a minimal learning curve, yet the capacity to scale (Pallets, n.d.-a). It has robust support documentation, an active development community, and is considered easier to implement than comparable frameworks such as Django. It uses the Werkzeug library to manage the WSGI details, including a debugger, request and response objects, a URL routing system, HTTP utilities, a test client for simulation, and a threaded WSGI server for

local development (Pallets, n.d.-c). It also includes the Jinja library, a Python template engine to support simple template design and inheritance. Key Python libraries include Shapely, SQLAlchemy, geoalchemy2, and Gdal.

SQLAlchemy: One of the key functionalities of the Apregoar application is spatial, temporal, and/or thematic filtering of story and instance attributes. To address this, the application utilizes the SQLAlchemy library which is a Python Structured Query Language (SQL) toolkit and object relational mapper (ORM). It uses a declarative mapping structure that simultaneously describes the real tables and their metadata in the PostgreSQL database, and defines these as Python object models that can be manipulated in the Flask framework. This is especially useful for filtering. For execution of simpler SQL queries, the psycopg2 Python library is used.

Apache: Managing the web app is an Apache HTTP server, selected for its ease of implementation and its recommended integration with Flask. Apache responds to content requests from web clients, serving static files and, via a WSGI protocol, communicates with the Python application, translating requests and responses.

GeoServer: For serving spatial features, GeoServer was selected for its vector data support, its documentation, support in literature (Bhattacharya & Painho, 2018; Jiang et al., 2020; Sami, 2019) and the researcher's prior experience. It implements Web Feature Service (WFS) and Web Map Service (WMS) protocols (among others) and is easily integrated with PostgreSQL/PostGIS. GeoServer includes the GeoServer application, which is hosted on an instance of Jetty, a Java web server developed by The Eclipse Foundation.

3.4.1.3 Client tier

OpenLayers: For mapping visualization in the front, OpenLayers and Mapbox were both considered. Mapbox ships a JS library with many of the operational, but not focal, functionality of the project. OpenLayers, however, was selected for its adherence to open source standards, continually developing feature set, community support, and future scalability. OpenLayers supports raster and vector data, layering and styling of distinct layers, and can support the manipulation of layers required for the filtering capabilities of the Apregoar project.

Stamen and OpenStreetMap: Additionally, Apregoar connects to map tiles of Stamen and OpenStreetMap, providing options for basemaps of various styles and informational levels. These contextualize the spatial information being entered into or queried from the Apregoar database. The inclusion of multiple sources/basemaps demonstrates the modularity of such layers. Stamen and OpenStreetMap were selected because of their licensing (map tiles under Creative Commons (CC) BY 3.0, data under ODbL), their compatibility with OpenLayers, PostGIS and GeoServer, and researcher familiarity.

GeoNames and Nominatim: As previously described (Section 3.3.3.1), Apregoar also connects to existing gazetteers for both contextualization and geotagging during geo-annotation. These libraries are called by the front-end in response to user search terms that provide matching results within the envelope of the Lisbon study area in JSON format.

Static files: Static files include all HyperText Markup Language (HTML) template files (including Jinja template engine), which are heavily augmented with client side JS scripts and styled with Cascading Style Sheets (CSS). Here, all of the programs, data, and visualizations coalesce in several application GUIs to provide the user with the informative and dynamic interface to access the geocorpora.

3.4.2 Application design

The web app is divided into several informational products (Section 3.1.2), each with its own interface(s) and user flows. Though all of these use the same underlying database, in practice they could be implemented as distinct tools. Therefore, each of these were grouped and implemented as separate Python files (named "views" in Flask, referencing related view functions), distinguishing the product functionalities from each other: Journalist views (generation of the geocorpora), Journal views (addition of context maps to articles), User views (Apregoar sign up, login, and log out), and Explore views (searching and filtering of the geocorpora).

Though data processing may occur on the front-end of the application (in the JS files and by various additional libraries), for efficiency the application should bias as much processing of the data to Python as possible.

3.4.2.1 Exploring stories

News stories are used by traditional readership to become better informed on what is happening in the world or in one's own backyard. If one is interested in entire swaths of areas, such as city wide, national, or international news, they may not care to discern between stories that happen in particular, sub-city locales. Other readers, however may be more selective and particularly interested in things that are occurring in areas relevant to them. They may be interested in news that happens near where they live, study, or work, and their commutes between them. They may also be interested in news that occurs near family members or friends, but not much for whatever happens between. These users may be interested in identifying news only in relevant locations that don't adhere to administrative boundaries. It may also happen that one has experienced an unplanned incident (such as saw a fire or an accident), or seen notice of a planned event and wants to understand more about it. To learn more, however, can be challenging if not already acquainted with the subject matter or a given name of the occurrence. As of yet, news platforms don't integrate spatial searching beyond the incorporation of keywords, or perhaps the choice of entire municipal areas. Likewise,

most temporal search is limited to the publication date. Stories are also leveraged for research purposes to understand things that have already happened in the more distant past for academic or operational understanding. In these cases, searching for such incidents may be particularly challenging as names of places may have evolved over time or have colloquial titles that the researcher is not privy to. In any of these scenarios, readers or researchers may be interested in the opportunity to define time intervals of the events of the news, as well as define their own area to return a map of associated stories (also filterable by their other ThAs) to better direct their browsing or searching experience. This kind of temporal and spatial searching has already been implemented into many of the applications used regularly all over the world, as described in Appendix A.2.

Therefore, the application should facilitate the definition of areas, themes, and times of interest by readers or researchers by using a mixture of pre-loaded options and free search options to define the geospatial, temporal, and thematic focus areas. The application should then return lists of the resulting stories with key summary data to assist a reader or researcher in determining their interest in the story without needing to open it. It should also return a map of the spatial distribution of the incident footprints.

When presenting results, the tool should reveal just enough details about each article to validate the interest of the readers (such as title and summaries, tags, etc.), from which users can access the original article situated in the digital platforms of the publishing agency. When searching for articles, the tool should include enough summary data to support complex searches. The underlying database should not include the full text of the original articles, as the web app is not a blog hosting site and does not need to function as such (Afzalan et al., 2017; Jiang et al., 2020). Rather, the Apregoar toolset should leverage the hosting capabilities of the original sources and link to these via URLs so that readers may continue their exploration experience there. This not only lessons the storage needs of the Apregoar toolset, but also creates a value proposition to the participating news agencies, such that they may receive additional user activity on their sites.

3.4.2.2 Contextualizing stories

It is understood that the integration of visualization into existing products is an important part of relevance and commercial value propositions (Meeks, 2019). As such, a key informational product that can immediately improve the user experience of local news readership is the integration of local maps associated with even a single story that can provide additional spatial context to a publication's readership.

Therefore, the application should simulate a news story with a context map allowing readers to visualize the distribution of the footprints of incidents associated with the chosen article. The map is interactive, such that users may scroll or zoom or use nearby

POIs on the basemap to orient themselves. Options to explore nearby stories should invite further user engagement.

3.4.2.3 Publishing stories

Publishers may be interested in harnessing the foreseen spatial benefits of geospatial news for their own research (identifying news deserts, drawing additional insights from the mapping of events etc.) and readership experience (contextualizing stories in space, providing better search mechanisms, etc.). Their articles may cover multiple events (instances) that happen in different locales at different times.

Therefore, the application should allow a journalist (or editor) to create stories and associate instances. Stories should have thematic attributes that apply to the article as a whole, such as newspaper, author, section, and tag, as a temporal attribute representing the date of publish. Instances should be composed of one or more geographic footprints, either reusing preexisting definitions or allowing the journalist to create their own. These instances should have their own attributes associated, such as a temporal definition of when the incident occurred, and textual descriptions of the place and time. The application should support journalist management of the stories and incidents, providing helpful list and map representations of the story and incident data.

3.4.2.4 User login

Registration to access the site is not necessary and browsing is completely unblocked. Contributing data, however, requires the definition of an account, which will verify the user as belonging to a particular institution or self-defining as un-affiliated. Likewise, personalization of settings (areas or themes of interest, saved searches, or other such planned steps) will require an account such that these values may be recalled at time of use. Otherwise, no personal demographic data is required, as it can inhibit participation from the community(Afzalan et al., 2017). The system should permit such profile creation and association of personalization and/or database management, as appropriate.

3.4.3 Relational data model

The publishers' entries will form a foundational database that will be accessed by each informational product and applied to the greater metropolitan area. This forms the beginning of a geo-annotated corpus of local news stories that can be used both by the Apregoar tool, as well as extracted by users for inclusion in other projects or studies.

Therefore, the relational data model should facilitate the functionality of the described user experiences, as well as permit functional management of the data, provision for future expected functionality, and consider possible uses of the data beyond its use

in the Apregoar toolkit (such as connection or download for external use). The relational database should store each story with incidents as a one-to-many relationship. Additional operational information, such as user profiles, should also be associated to ensure appropriate recall of stories by publisher, by publication, or by any personal preferences or history users may elect to set.

3.5 Validation

To rigorously validate this work, the Apregoar tool should be implemented into the workflow of several digital newspapers serving local communities in Lisbon (as defined in this project, any area falling within either the District of Lisbon or the Metropolitan Area of Lisbon) during a specific timeframe, and compared against other automatically georeferencing tools (such as those evaluated in (Rivera et al., 2020) on the same corpora. By comparing the results, one could empirically determine if there is in fact an improvement in reference accuracy and precision. The system is expected to outperform automatic, large scale, international endeavors (see Section 2.3.3), as these do not attempt hyperlocal granularity.

To address the value of the tool, the Apregoar tool could again be layered into the workflow of at least one digital newspaper, where it could be A/B tested (in which readers would be randomly segmented into groups, one of which would be able to access the Apregoar tools and one would not) to determine if (via survey) customer satisfaction increased or (via time spent on website) customer engagement increased with access to the tools.

Both of these validation techniques are planned future steps for the project, once it is transformed into a functional plugin that can be implemented into the back office systems already utilized by local newspapers (see Section6).

Until such a point is reached, validation focuses on accuracy of recall of story details, incident details, and geometries, as well as how well the Apregoar web app can accommodate the details of the geocorpora.

4

Results

The following sections describe the current status of each of the planned/developed tools of the Apreogar project.

As the full code for the Apregoar project is available via GitHub, the following describes some of the high-level and more interesting design choices.

For immediate next steps for each tool/product, see Section 6.1.

4.1 Explore tool

While using the explore tool, readers or researchers can view articles in list or map format. Additionally, they may filter the geocorpora using the following options:

- Spatial filters: a) Where the story contents are happening.
- Temporal filters: a) When the story contents are happening; and b) When the story was published.
- Thematic filters: a) Type of gazetteer: custom vs. existing; b) type of temporal definition: persistent vs. date range vs. time range; and c) Attributes of stories: source, section, tag(s), author(s).
- Text searches: a) place names; b) Descriptions of time and place of instances; and c) Descriptions of stories.

Upon page load, the web app back-end extracts all distinct values of ThAs related to story and instance, minimum/maximum datetimes of instances and minimum/maximum dates of publish and presents these as options or limits (as appropriate) to various fields in a filtering form in the front-end. It also extracts the date range of the 25 most recently published stories to as the default filters to show upon load. See the first part of Figure 4.3: ”Load Explore” to ”Explore Map” blocks.

To interact with the database, users open a form in which they may define all search criteria. Though automatic update was considered (”onclick” event triggered searching), the nature of the complex queries and WFS calls incurred a state of constant

loading. Therefore, the form method was implemented encourage more intentionally, fully formed search queries.

Users may draw areas of interest using a freehand input style, in which the motion of the user's fluid, drawn input is approximated by the application with many vertices and connecting line segments, which as a faster method is expected to better suite the needs of readers or researchers. Users may further refine results be selecting if the results should be completely contained by the drawn area(s), partially or completely contained the area(s), partially contained by the area(s), or excluded by the area(s). Though some of these search terms may seem redundant, this allows one to filter out stories that apply to much larger swaths of the area (such as continental Portugal, districts, or municipalities). Users may also search administrative areas, which will return not only the instances to which these administrative gazetteer records were assigned, but also all of the custom gazetteer areas overlapping these (see additional details about this automated spatial enrichment in Section 4.3). See Figure 4.1 for a snapshot of the GUI of the Explore filter tool, and Figure 4.2 for a snapshot of interactions with the search results (larger image and additional web app screenshots can be found in Appendix F.2).

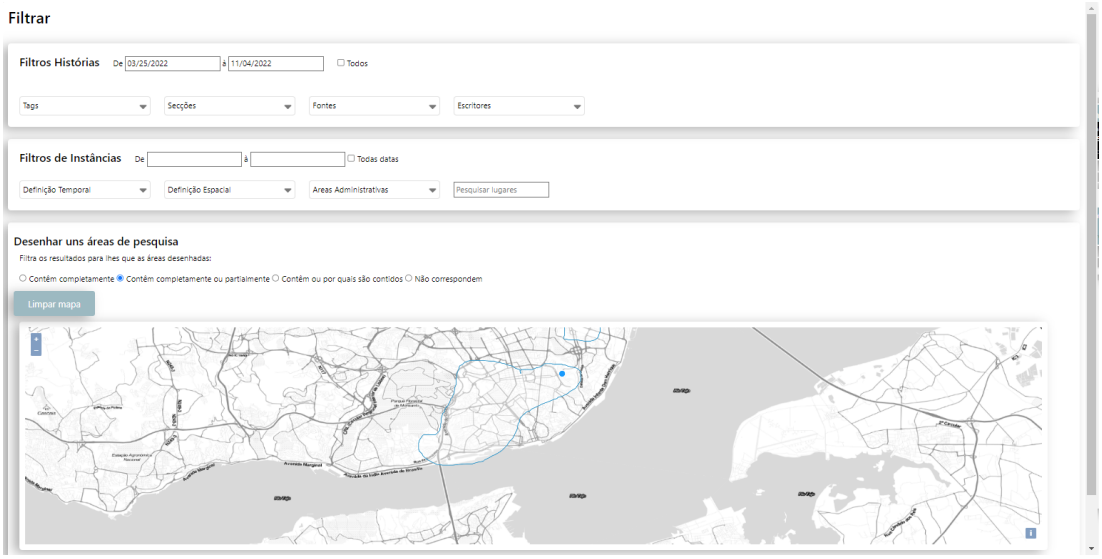

Figure 4.1: GUI of filtering functionality in the Explore tool

The default or user-defined filter values (including multi-select checkboxes, search terms, date and time ranges, and drawn areas) are extracted from the form and packaged as geographic Javascript object notation (GeoJSON) to the back-end. There, the Python script, using a sequence of if statements, determines which filters have been applied and builds a modular query in SQLAlchemy returning all of the applicable records. Multi-select options are treated as OR logic (the record values could be one or more of the selected options), and the combination of multiple options is treated as AND logic (the values must adhere to all of the conditions specified). This logic

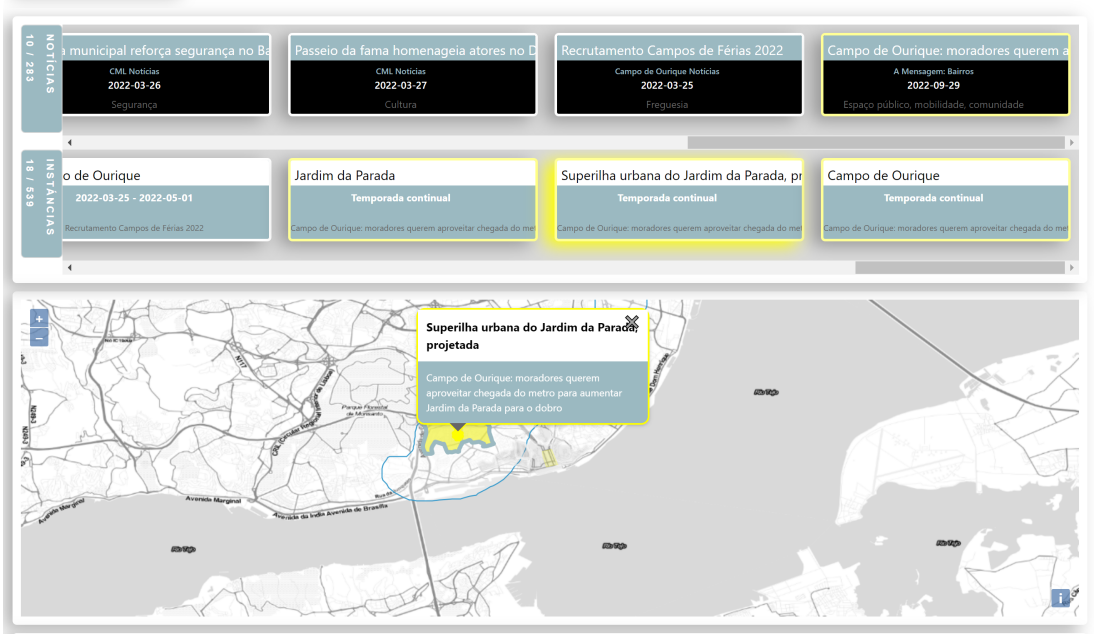

Figure 4.2: Highlighting of story and instance selection in the Explore tool

system was selected as it is common and expected functionality in many user-focused applications and websites. The information is saved as a response object with additional nested objects representing both story and instance focused objects, and lists of instance IDs. Though redundant, this facilitates the organization and exploration of the data, without much data-processing necessary on the front-end. See Figure 4.3: ”Explore Map” to ”Prepare results and package as JSON” blocks.

On the front-end, the JSON response is received and a WFS call is made to GeoServer to load the relevant instance features and update a map of the results, centering and zooming to the contents. WFS features were chosen instead of WMS maps, as these can be more easily manipulated and styled by the front-end application, though it sacrifices the speed of the simple WMS map. Lists of resulting stories and resulting instances are loaded and exposed. See Figure 4.3: ”Parse JSON response” to ”Explore Map” blocks.

The styling of the return features uses semi-transparent fills and bold outlines such that the areas of each footprint are clear, and areas with more stories appear darker. Outlines default to black, though these are updates as users interact with them.

Interactions with the Apregoar Explore GUI permit the visualization of additional details, as well as the highlighting of relevant map features or list items as they are selected (Figure 4.4). For example: selection of a story will automatically highlight all of its instances in the instance list and on the map, and lowlight (highlight to a lower degree) all other loaded instances. Likewise, a selection of an instance from the list will automatically highlight the story in the story list, and all other instances related to it, while brightlighting (highlighting at the highest level) the selected instance and loading

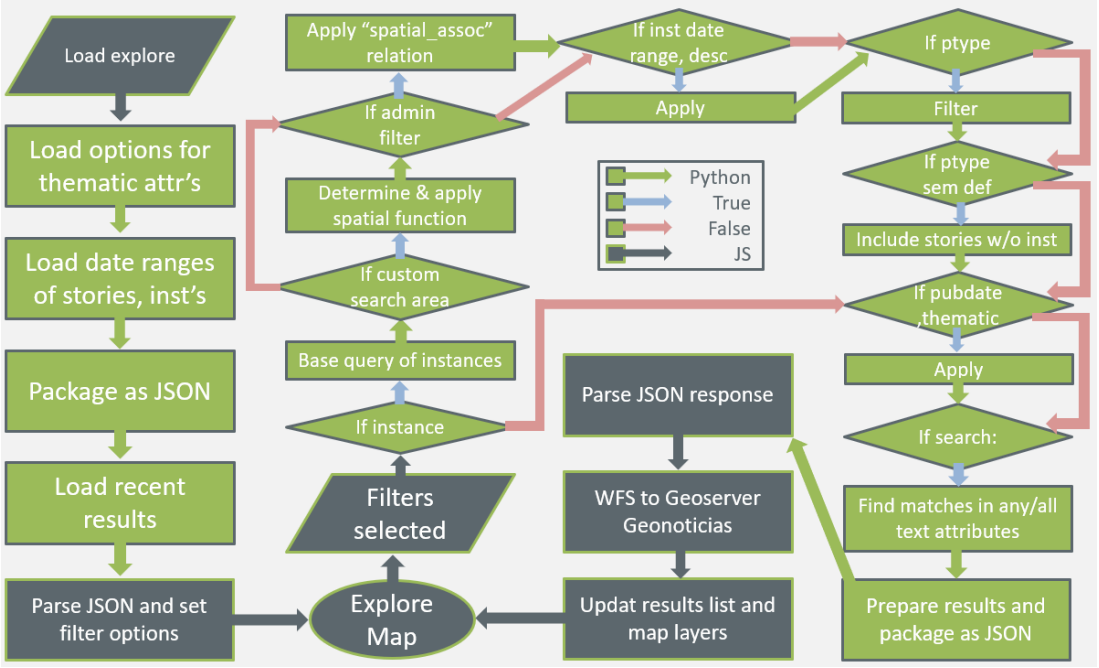

Figure 4.3: Flow diagram of the load and filter processes in the Apregoar Explore tool

a map popup with some details, and lowlighting the unrelated loaded instances. Finally, a click on any point of the map will perform a WFS call to return and highlight all features at that point, with a scrollable popup that permits navigation of these results and adjust the highlighting of these, while lowlighting all unrelated features. Any of these interactions load the relevant story and its instances to display an overview of the loaded data, with an option to follow a link to the original full story. This process is meant to lead users to the original story/ies that match the user's thematic, spatial, and temporal interests.

See Appendix F.2 for an example of this Explore tool including screenshots of the filtering, highlighted results, and story details GUIs.

4.2 Context tool

Readers can access articles with context maps displaying the footprints of the article incidents and a panel from which they can explore nearby stories.

When a story is selected (such as from the Explore GUI), the Apregoar back-end extracts all story and instance details of the chosen story and extracts the related gazetteer IDs. Based on previous auto-association of custom gazetteer entries to overlapping administrative polygons during the publishing phase (see Section 3.4.2.3), it loads the names and IDs of these spatially relevant areas. These are prepared as JSON data and sent to the front-end. See Figure 4.6, blocks "Selected story" to "História".

The front-end then parses this data and attempts to load an iframe (short for "inline

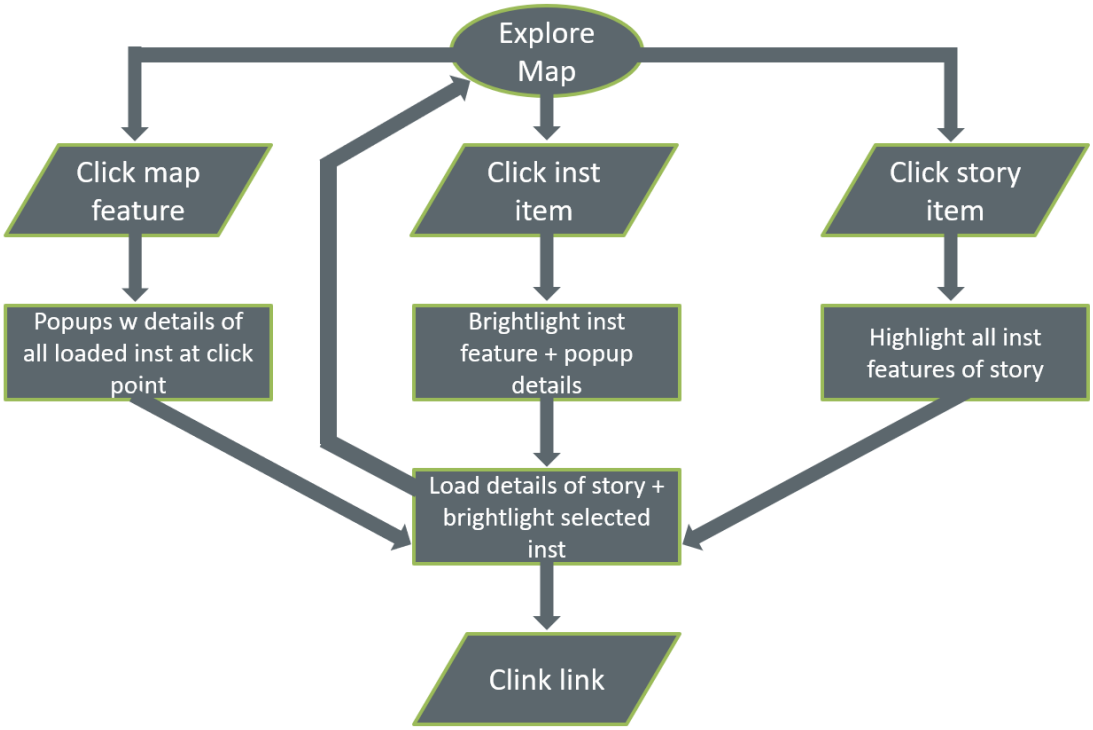

Figure 4.4: Flow diagram of the navigation processes in the Apregoar Explore tool

frame", in which other online content is embedded into the page) of the original article via its URL. If the process times out, the page defaults to displays the pre-loaded story details. The front-end communicates with the GeoServer to load WFS geometry. These are displayed at the bottom of the page a contextual map, zoomed to the related footprints. Like the Explore GUI, these support interactivity to show additional details about the incidents present at a particular point. A panel just the left of the map shows names of nearby areas that may also be of interest to the reader. See Figure 4.6, blocks "História" to "Similar area".

Readers may select a nearby location, leading towards a simulated, agency white-labeled (showing only the agency's articles, sections, journalists, etc. such that it is integrated into the browsing experience) explore page of content in the selected area that have been published by the same news outlet. See Figure 4.6, blocks "Similar area" to "Identify publication". The tool provides an agency specific GUI that mimics the functionality of the general Explore tool. Note that the blocks in light green and grey represent the next planned steps for integration into the next revision of the tool.

Figure 4.5 shows a screen shot of the Context tool applied (see Appendix F.3 for an more information and a larger version of the Context GUI image).

Figure 4.5: GUI of simulated original article with contextualization map

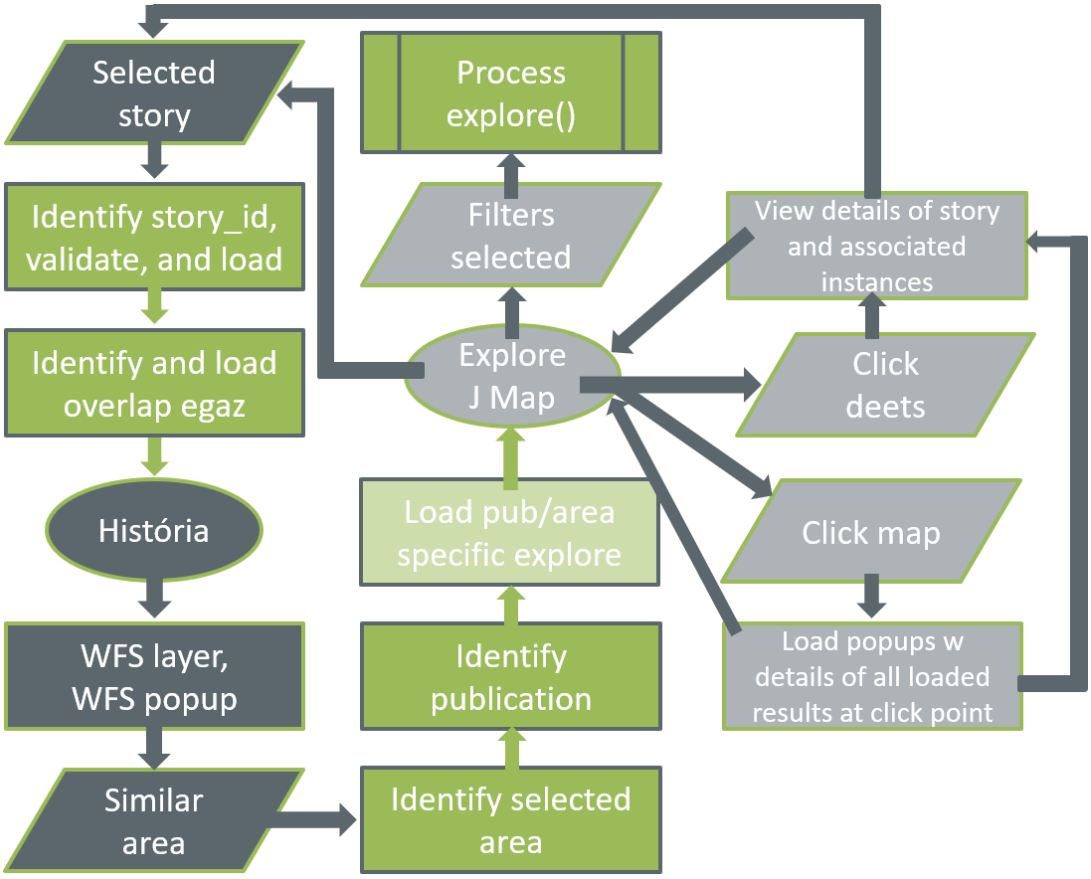

Figure 4.6: Flow diagram of the contextualize processes in the Apregoar Context tool

4.3 Publish tool

This process simulates the publishing process used by digital news agencies during the creation and publication of new articles (often referred to as blog entries in their back office tools). Ultimately, this feature should be available as a plugin to existing news agency back office tools so facilitate the user experience. However, as a development endeavor with a new set of tools for integration with WordPress, that is outside of this project's scope as a proof of concept. Instead, the journalist must duplicate data input by copying certain values from their already digitally published articles.

A user login is required to access the publishing tool. The associated user id is stored as a global variable, and this value is automatically associated to any elements (stories, instances, gazetteer entries) published the user so that these can be later associated to the appropriate user and agency, and viewed or managed from their perspective dashboards.

A journalist may "publish" stories by creating a new story entry, and inputting key descriptive values – title, author(s), publish date, and URL – copied from the existing articles. Optionally, they may associate a section, multiple tags, and a summary. These values can also be used to filter the stories in various informational products.

Upon saving, the Apregoar front-end validates that all key values have valid entries. The data is then sent via JSON data to the back-end where it is extracted and the URL field is checked against the existing URLs in the stories table in the database to confirm that the story is unique. Uniqueness is determined by URL, as it is possible that different stories (unrelated themes or coverage of the same issue from a different source) may have identical names.

The back-end then creates a new record in the stories table and a unique story ID is associated. The story attributes are then distributed into tables in the database. See Figure 4.7, blocks "Create" to "Distribute As into tables in DB". Upon completion, a webpage displaying a summary of the saved story is rendered.

At this point, a journalists may elect to associate one or more instances, these describing the where and when of the events in the story's contents. On the localization page, journalists must associate a name, one or more polygons and a timeframe to each instance, and may input additional information in the description sections of place and timeframe for further context.

When associating place, journalist may use pre-existing places or design new entries. To peruse existing options, journalists may load lists of a) personally defined toponyms; b) agency defined toponyms; c) all Apregoar toponyms; d) current freguesias; e) current concelhos; f) current administrative groups; g) green spaces; or h) archived administrative boundaries, which are separated to support usability (they are pre-organized by type) and efficiency (sub-lists are loaded as needed instead of loading all gazetteer records). This allows easy reuse of georeferences, and also supports author- and agency-specific local lexicons. When any of these gazetteers are selected,

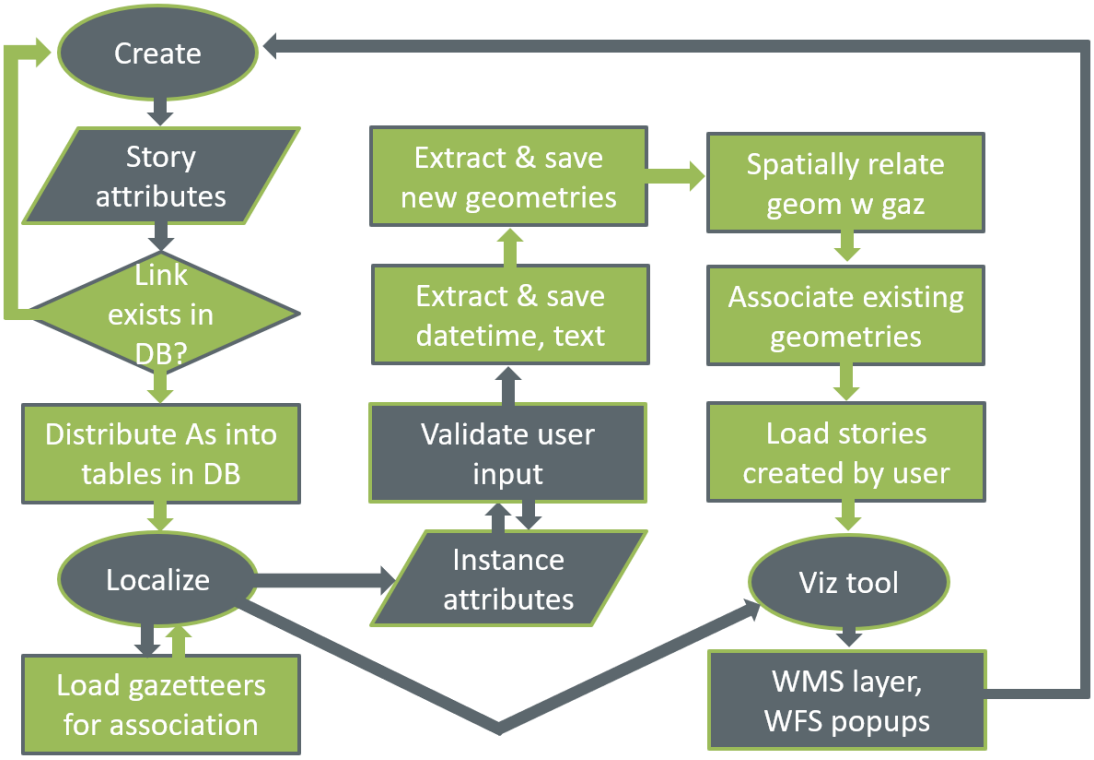

Figure 4.7: Flow diagram of the publish processes in the Apregoar Publish tool

the back-end will retrieve the matching values from the gazetteers and send them to the front-end. In the event that a journalist is not sure in which list a toponym may appear, they may search all existing areas, which will return the matching values organized into the above mentioned lists, as well as search the OSM API for any potential matches. The OSM API may also be searched individually. Upon return of the matching place names from their respective gazetteers, users may select one or more of these and preview them on the map.

Journalists may also elect to design a new footprint if no pre-existing polygons match that of the incident in question. The journalist may define one or more polygons on the map, rendered in a different color to distinguish this from any previewed gazetteer entries. The custom polygons utilize a snap design that requires the user to place multiple points (vertices), which are automatically connected with line segments by the program. This functionality (versus freehand) was chosen for the geometry definition as it allows for deliberate placement of vertices and longer, straight lines between them, which is expected to support user definition as often they will be tracing features on the basemap such as buildings and roads.

Any designed geometries and previewed gazetteers will be associated to the incident upon saving (mixes of gazetteers and custom designs are valid). At least one geometry (custom or existing) is required to save.

When associating a timeframe, journalists may choose between all day events (date

format), hour specific events (datetime format), or persistent events (no dates associated). Persistent events are appropriate in situations in which a place is described, but no specific timeframe is appropriate. User updates of the timeframe data formats will automatically update the required fields.

Upon saving, the front-end will validate the user input then extract bundle the values into a JSON format. In the backend, any new geometry entries will be saved as new custom gazetteer entries, and any geometries will be related to the incident. These custom geometries will also undergo automated spatial enrichment, which will determine to which administrative boundaries they belong (such as freguesia or municipality) so that they can be recalled with these groups more efficiently. Upon returning to the article summary page, the back-end loads all story data associated with the story id, as well as all of the associated instances. These are displayed on an article summary page, where the front-end retrieves the WMS map layers of all instances and displays them on a story map. Journalists may then return to a management page in which all of their stories and instances are loaded for review and further management. See Figure 4.7, blocks "Localize" to "WMS layer, WFS popups".

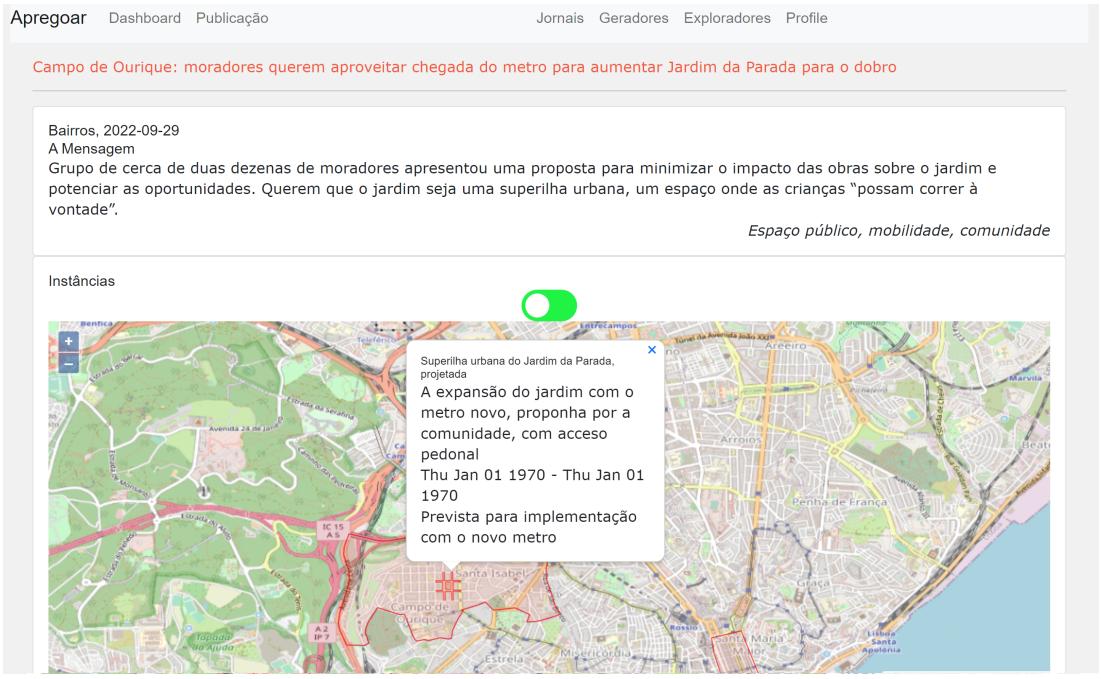

Figure 4.8: GUI of a "published" article from the agency's backoffice

Figure 4.8 shows a screenshot of the Publisher tool GUI (see Appendix F.1 for an extended example and larger image).

4.4 Spatial news database

The relational data model underpins all of the functionality of the various informational products, both implemented and planned.

Core to the design is a table representing the geocorpora of stories, each record of which is associated to a single published article. Each story must have a title, date of publish, a unique web link, and publication (newspaper, website, or magazine name), while other descriptive attributes (summary, section, tags, author) are not required. Automatically, the ID of the user publishing the story (u_id) and timestamps of creation and last edit are assigned upon creation of each story record. Various other tables, including publications, tags, authors, and sections are related to the stories via join tables to support simple filtering while avoiding many-to-many relationships.

Instances may be optionally associated to a story. Each story may have zero, one, or several instances associated. An instance may only be associated with one story. Each instance describes a place (which may be a single or multiple polygons), and ascribes a timeframe to it, which is defined by datetimes "t_begin" and "t_end". Additionally, these are assigned types "t_type" indicating that the timeframes are all day (i.e., that the times associated with the timeframes should be ignored: "allday_y"), not all day (the associated times are important: "allday_n"), or persistent (which should ignore start and end types, "allday_p"). Additionally, each instance must have a user assigned name ("p_name"). Optional descriptions of time and location are recommended for inclusion at the journalist's discretion. Automatically, the ID of the journalist and creation/edit timestamps generated at the moment of saving are associated.

The place describing where an instance occurs is defined by associating gazetteer entries. Each instance may have multiple spatial definitions associated with it, and each gazetteer feature may be associated to multiple instances. Further, multiple types of gazetteers may be associated with a single instance. The gazetteer of official existing features hosted locally in the PostgreSQL database (nicknamed "Egazetteer") includes existing features of area types from external sources (such as the CAOPs or CML Geodados). The local instances of remotely connected gazetteer features ("Ngazetteer") is a growing repository of already associated features from the OSM Nominatim database that are accessed via their API, then saved locally to the PostgreSQL database so that their key attributes may be recalled more quickly by the Geonoticias query (see Section 4.5). The user gazetteer (or "Ugazetteer") is a growing repository of user defined places during instance assignment that may be associated to other articles, with attributes defined at their creation. Ugazetteer features are spatially related to Egazetteer features upon addition of new features to either gazetteer to fulfil the automated spatial enrichment functionality. These are saved into a spatial_assoc table that includes the Ugazetteer ID, Egazetteer ID, and their spatial relation.

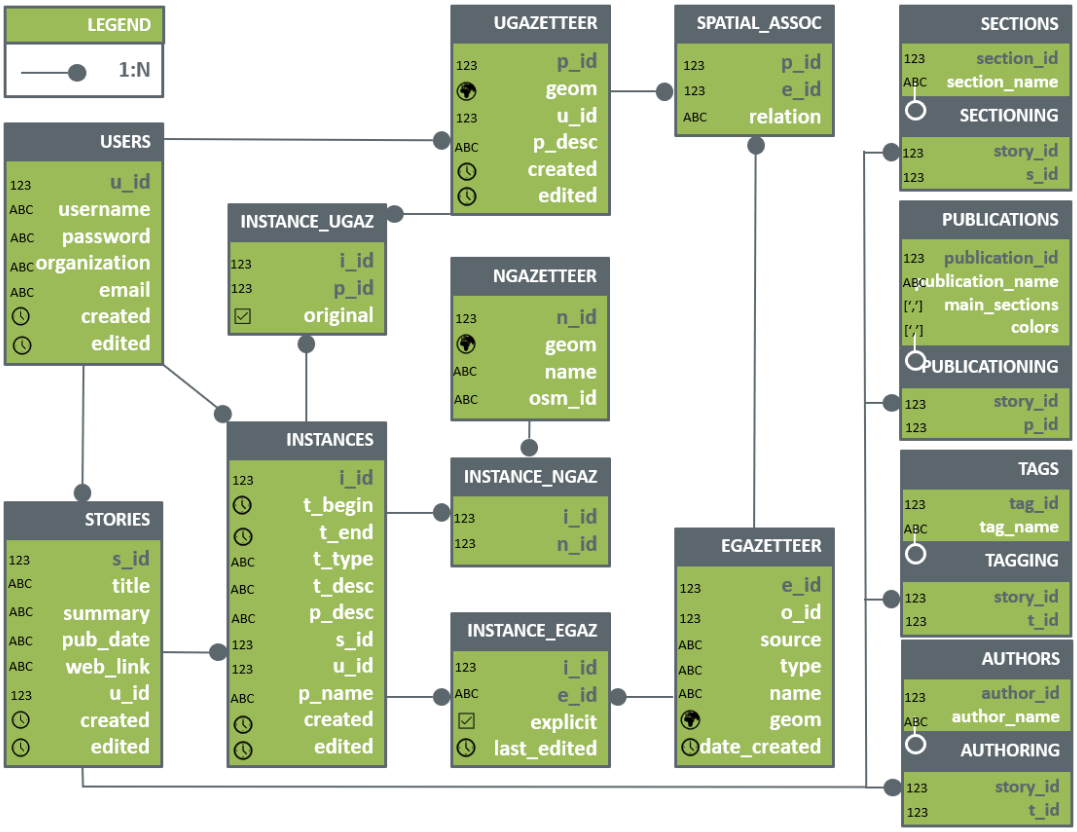

Figure 4.9: Apregoar data model

4.5 Queries

Several complex queries (also called 'views' within the local vocabulary) were created and named in PostgreSQL to facilitate querying the database from the Apregoar web app.

The most pertinent of these is the Geonoticias query (Figure 4.10), which results in a virtual table in which each record represents an instance with its related story attributes, summaries of the associated gazatteer attributes, and a single field including all instance specific geometries (a collection of egazetteer, ugazetteer, and ngazetteer features). Each story, therefore, may be represented multiple times (if associated with multiple instances). Stories without any associated instances are included, with null values in place of instance related fields. This virtual table is accessible through database connections from the web app, without the need to reconstruct the same complex query. GeoServer also connects to this query, so that the features can be served in WMS or WFS formats to the GUI, where the end users may view or manipulate the results.

See this and additional queries in Appendix G.

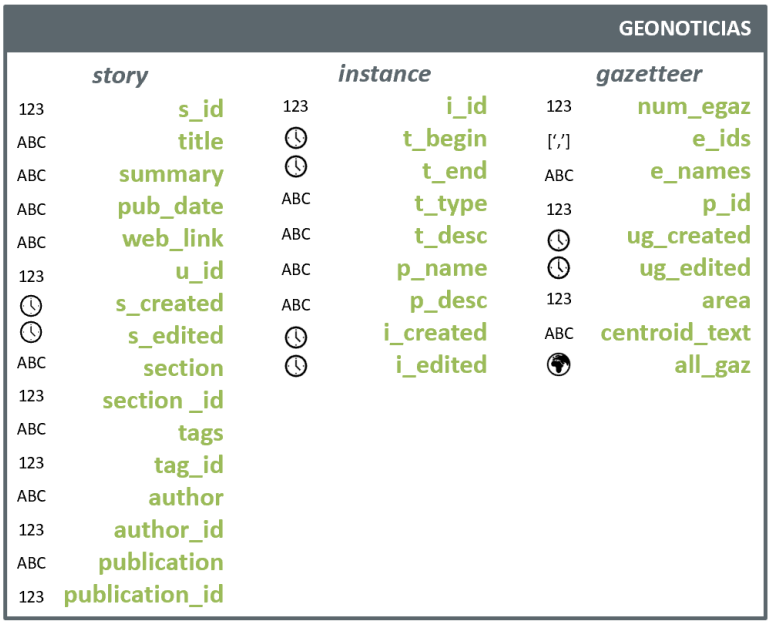

Figure 4.10: Geonoticias query

4.6 Validation

The Apregoar web app was developed and tested using Google Chrome browser. It is currently locally hosted, but with plans to be accessible after additional testing and refinement.

Though testing of these hypotheses through rigorous comparison to the status quo (traditional online news sources without a spatial element) and emerging products performing automatic extraction of place are not included in this endeavor, the resulting tools should provide a basis from which future projects may develop and evaluate.

Validation at this stage is based on visual inspection for accuracy of geometry of the geocorpora. At this stage, since the georeferencing is not being associated by either the original journalist or agency, it is likely that the geometries are not exactly adherent to the appropriate locations. For this reason, rigorous testing should be done by these professionals to ensure accuracy, ease of entry, and value. This is the next step in the development process.

It should be noted that the original gazetteer for POIs was GeoNames, as per myriad of recommendations from the literature. However, GeoNames does not include many records for sub-city level toponyms. Therefore, only after the geocorpora was introduced to the Apregoar database, was the OSM Nominatim connection implemented. Some incident locations that were challenging to map during introduction via the Publication tool are more easily associated using the Nominatim gazetteer. However, some locations are still nebulous, as is described below.

Two main challenges were encountered, both of which are demonstrated by the attempt to introduce the A história das placas toponímicas de Lisboa africana à espera

de ser postas por falta de verba story into the Apregoar database.

The first challenge refers to the tendency to summarize data in articles. The article references multiple placards inserted throughout the city, not all of which included reference locations in the text. If associated by the actual author of the piece, perhaps all of the locations could be precisely referenced (and therefore, visited by readership) without needing to textually describe them in the article contents. As it is written, it is impossible to know where one can find the placards throughout the city.

The second reiterates many of the ambiguity challenges described in detail in both the literature review and methodology. Often, the places mentioned in articles did not reference administrative boundaries or POIs from the OSM Nominatim API, and are therefore challenging to map without more familiarity of the story contents. In the article referenced above, "Bairro de Mocambo" (and variations thereof) does not appear in the gazetteers associated with Apregoar, nor can it be easily defined using textual references in internet documents searched online. Therefore, the researcher used what information was available to approximate these areas.

Moreover, the lack of formal definition of neighborhoods continues to contribute to ambiguity. Efforts to contact CML requesting any sort of these sub-administrative definitions has resulted in references to their online tools at Lisboa Aberta and GeoDados. While both of these are excellent resources provided by the city, they are holes within their coverage. For example: some of the GeoDados layers reference specific neighborhoods (such as the Zonas 30 Mobility Map layer that identifies six neighborhoods or Limite de Planos de Urbanização with 10 areas designated), though these are far from comprehensive of the entire city, let alone the entire study area.

5

Analysis

5.1 Impact

This project provides the basis of a web app, soon to be released freely and openly accessible, as well as its development code, which shall be openly licensed for further or related future development by any individual or organization.

Apregoar is a POC of a geo-annotation tool that can be used directly by journalists or editors to provide the basis of a growing geocorpora, from which further information can be extracted. These data can be used to contextualize a story for the individual, to inform community action, or as an additional layer for geointelligence applications. It also sets the foundation for commercial value to the news organizations that adopt it, as this kind of functionality may attract, engage, and better inform readers, as well as provide additional insights to the agency's coverage.

Unlike other third party georeferencing tools, these are associated by the journalist to provide the best description of events in time and space, whether or not they adhere to existing administrative boundaries, POIs, or other gazetteer entries, and without needing to divine the intention of the author. The next version of tools will also provide a means for adjustment of these geospatial and temporal definitions even after publishing.

The Apregoar toolset does not require any sort of formal training or previous experience with GIS to successfully define incident geolocations, nor to extract meaning from the presented map graphics, supporting neogeography.

5.1.1 Use cases

The following are examples of potential use cases of the Apregoar toolset and data products applied to external data products:

1. Lisbon, PT: A website to inform citizens on the distribution of COVID within the city limits. A development team obtains i) locations of hospitals and nursing homes, locations of vital businesses (markets, pharmacies, gas stations, etc.),

and boundaries of freguesias from Lisboa Aberta (open data portal managed by CML); ii) news stories related to COVID (results from the toolset filtered to Lisbon Municipality boundary and 'COVID' tags); and iii) aggregation (by freguesia) of active COVID cases as well as identified hotspot areas from the Ministry of Health. Users are able to monitor their locations of interest (home, work, play, family, and friends) in terms of relative cases compared to the rest of the city and adjust their behavior in those areas accordingly. Trips to vital services can be planned more effectively. City officials may also identify areas of poor coverage for vital services and temporarily support those areas with access to walkable points of pickup for food and other necessities.

2. Campolide, Lisboa, PT: An environmental task force for the freguesia of Campolide seeks to better understand the ecological situation in their community. Their team incorporates i) news stories within the Campolide boundary filtered to the 'ECO' tag, all mentions of keywords 'JARDIM' (garden), 'ESPACO VERDE' (green space), 'ECOSYSTEMA' (ecosystem), 'QUALIDADE DA VIDA' (quality of life), and temporal range of the last year (17 Nov 2019 to 16 Nov 2020); ii) location of green spaces within the freguesia from Lisboa Aberta; iii) pollution and trash data from other sources; iv) citizen interview results. The task force can then perform the necessary statistical analysis to determine areas requiring intervention, or identify those areas to highlight in upcoming reports on successful green spaces within the community, or to learn about new grassroots initiatives to support.

3. Berlin, Germany: A research team is searching for statistics on the effect of the Spanish flu in Lisbon, Portugal. They are not well-versed with the study area nor with the Portuguese language. They use the toolset to segregate their own study areas (not adhering to current or previous administrative boundaries of freguesias) within the city limits by drawing polygons over the region and setting a temporal content filter to January 1918 to December 1920, with tags 'DOENTE' and 'GRIPE ESPANHOLA', and filter this again to publish dates of each pre-2020 and 2020. They incorporate this into their own study methods. They are able to visualize the distribution of results of each query on a map, adding to their understanding of the coverage of each topic.

5.1.2 Spatial awareness

From the outside looking in: By employing spatial searching options to a geo-referenced corpora of events in the news, end-users are able to use not only textual references to search for relevant articles, but also draw on a map the physical area of interest. Official designations of area are always changing – administrative boundaries are redesigned every so often (such as in the administrative reorganization of 2013 in continental

Portugal, in which the existing 284 freguesias were reworked into the current 170) – and areas that have historically anchored textual references may be renamed, evolve or move over time. Names may differ between cultures, generations, and or languages. By referencing geocoordinates, much of the reference ambiguity is sidestepped and all events with footprints within, overlapping, or outside of (as per the user's preferences) a drawn area of interest can be returned without requiring specific local knowledge of place names (see Figure F.2).

From the inside looking in: As technology has broadened our horizons to include more and more territory, some of our focus may have been diverted from what is physically happening on our block (Painho & Pina, 2013). While we laud the applications that expose us to new ideas, information, and media outside of our physical sphere, recommendation algorithms may evaluate our tastes to fall into pre-defined tracks and hide all other content from view. By exposing users to those things that have or are plan to happen near them, it may provide them with a new means to reconnect with their surroundings, accessing layers of the city of which they may not have been aware.

5.1.3 Public participation

The tool can be considered a public participation geographic information system (PPGIS) system (see Appendix B.2) in that those assigning spatial definitions will not be trained GIS users. Rather, they will mostly be composed of newspaper journalists or editors, and communication departments of public institutions. An even more broad definition of "public" is possible though, as non-institutionally affiliated users are also welcome to contribute their own georeferenced stories. It can also be considered a PPGIS as it will require the assignment of place on both objective place definitions (addresses or existing boundaries) as well as custom definitions (areas that don't conform to administrative boundaries, understood point of interest, or incorporate multiple areas) (Brown & Pullar, 2012). The visualization of this data is then available to any user for further exploration at will (Figure F.2).

Its greatest value, however, is as an input to other PPGIS, helping to form a general spatial understanding of the community by providing additional insights to citizens about the areas that affect them – where they live, work, play, or have an interest (Evans-Cowley & Hollander, 2010). As a dynamic evolution of placemaking within a local community, the tool aggregates and georeferences the city's own public information onto a navigable map, but can invite and accommodate commercial information (online publications) as well as private citizens micro-blogs to gain further citizen feeling or commentary about their physical surroundings.

5.1.4 Geointelligence

Because journalists have georeferrenced the articles themselves, the resulting high-confidence datasets are a potential input for GEOINT applications (Teitler et al.,

2008). The tool should eventually support situational awareness via dynamic dashboards (Varanda, 2020), either directly incorporated into the tool or feeding into other more developed GEOINT systems (Datta, 2018).

5.1.5 Data literacy

Everyday users are becoming increasingly exposed to and interested in geopolitics (Granger, 2020) as well as data visualization tools. Both of these have been apparent during the COVID-19 pandemic news coverage that inherently has a spatial component, which has been the study of many great minds as future scenarios are predicted and immediate preventative measures have been enacted at every level of community, from the parish to the international regional level (Granger, 2020). The layering of geonews with other spatial data in a geoportal may support any type of user to better identify the data they require and notice trends among seemingly disparate data records (Jiang et al., 2020).

5.1.6 Psychographics

In addition to colloquial data (social media), geolocation of official reports can contribute to both how an event is felt within and outside of a community. The differences between styles of reporting can be evaluated from place to place, as well as the relative importance given to a place by various communities (Datta, 2018). When paired with other data sources such as internet of things (IOT) or VGI, the impact of any given event may be better understood for its influence on others, such as reporting an election's impact on societal activity (Bhattacharya & Painho, 2018). It may also be possible to better detangle co-occurring events and the possible impacts that they have on each other, such that the impact of one event isn't misattributed to another simultaneous one.

5.2 Concerns

5.2.1 Confirmation of biases

In addition to some of the potential valuable outputs of the system, there are, of course, concerns keeping page. Throughout the development process, one should consider: What kinds of potential themes may emerge (accurately or otherwise) in relation to place in Lisbon? and How should these be dealt with? Though this project doesn't any have any external visibility yet, it is meant to identify trends, both actual (events occurring in time and space) and contrived (the coverage of events may not track their occurrences, due to biases intended or unrecognized at any level of the agency story selection process) (Baker et al., 2019). These trends may help to identify where

interventions are warranted in a city, or may confirm stereotypes about places existent in the communities.

5.2.2 Consistent application

The value of the Apregoar informational products is reliant on consistent and appropriate feeding by those publishing stories. On a small scale, each individual institution may create value by providing an improved user experience to their readers via a white-labeled integration, as well as added spatial and temporal dimensions to their own internal analytics. However, for large scale value to be extracted from the Apregoar database, ideally many different agencies would contribute to the growing geocorpora to produce more complete and diverse coverage of an area. This, of course, would require a sufficient value proposition to potential publishing user, such that they would be motivated to participate to improve objective metrics (new customers, customer retention, engagement, perceived transparency, efficiency, new products, etc.).

5.3 Sustainability

The project is a foundation for future development in the geospatial and temporal distribution of news story contents. The proof of concept should demonstrate the value of such filtering and may be built upon in one or more of the following ways:

1. as a free tool;
2. as the base of a new online news journal product;
3. incorporated into existing online databases to incorporate the temporal spatial dimension into and enhance their own thematic tools; or
4. to be incorporated into municipalities as a public participation platform / community empowerment tool to better understand incidents that are spatially relevant.

At minimum, its documentation and codebase will be available under an open license from which anyone may develop in the future.

This last option is especially interesting if future planned events and city data are layered in. It is also the direction of most interest to the author, and future efforts may involve collaboration with one or more cities to design a public participation tool.

A proximal step is to develop a business plan for further exploration of the implementation of the concept in select partners in the Lisbon area: such as incorporation into the freguesias or municipalities, and/or local newspapers.

Section 6 describes the immediate next steps and roadmap for future development.

6

Conclusion

News is a distribution of information to humans which in turn affects their actions. We have manners of measuring the physical world via sensors, and now are layering in emotional responses. The news is a feedback loop that is at once an input to human decision making and a report of some of those outputs.

This work describes the development of a POC, open source (OS) and free software (FS) web app that supports the visualization of the spatial distribution of news story contents ("incidents"), as well as filtering mechanisms for improved temporal, spatial, and thematic investigation of news articles. It demonstrates how the geospatial element provides an additional dimension of understanding, allowing users to better contextualize news stories, search repositories, or monitor spatial/temporal trends at a community level (within a city), especially when combined with other layers of information in application specific geoportals.

In addition to the aforementioned improvements of user experience for the public (readers, researchers, and monitors), it is also expected to support publishers via the inference of new insights from their existing internal data, such as the illumination of under- or over-reporting of areas by theme for better investigative coverage. Ideally, this functionality could be expanded to integrate multiple sources, as well as the incorporation of planned events and/or resources to provide a more comprehensive understanding of one's surroundings in both the planned future and transpired past.

6.1 Next steps

Immediate next steps are described below.

Exploring interface:

- Create a new reader login portal to support the personalization of reader search preferences (localizations of interest and tags to support notifications of new stories of interest) and saved queries (of complex queries perhaps most useful for research applications) associated with their profile. This supports the answering of the "what happened here?" question. This functionality will require a new

table in the database (user_reader) in the database and a query summarizing the associated preferences and queries.

- Improve GUIs of search filters such that they are more intuitive. For example: implement bubbles of activated features.
- Incorporate hyperlocality searches using the current geolocation and radius as the search criteria. This answers the question: "what is happening near me?" more automatically than manually inputting current location.

Contextualization interface:

- Implement popups to mimic the functionality of the Explore GUI for a more consistent user experience across tools.

Publisher interface:

- Implement a main photo of each story to ease the browsing experience, as is customary in many similar systems (see Appendix A.2).
- Include "verified" tags to stories if they are added by the publishing organization (versus added by an unassociated third party) to boost confidence in accuracy of search results.
- Include an "GeoInstrucções" section to help guide journalists in the best practices for inputting spatial data for this purpose.
- Create a method for journalists to import existing spatial definitions (such as polygons provided by their sources or existent on open data platforms, including GeoJSON or shapefile formats, or connections via APIs) to reduce the labor of manual definitions.
- Rework the input of agency specific attributes ("Author", "Newspaper", "Sections") so that these can be selected from agency specific dropdowns to improve efficiency and user experience.
- Restyle the GUI. This was developed early in the process and therefore should be reworked for consistency throughout the toolkit.
- Allow journalists to not only create, view, and delete stories and instances, but also to edit them.

Spatial news database:

- Add additional article records from a variety of sources to continue to test the current structure and increase coverage.

Queries:

- Create additional queries representing reader personalization in their profile.

6.2 Development roadmap

6.2.1 Implement a monitoring interface

Though not yet implemented, monitoring dashboards are one of the most interesting potential applications of such georeferenced data. These provide the opportunity to layer GEOINT into the platform, supporting decision makers to take action based on the produced findings.

Users of the monitoring dashboard may leverage a variety of statistical techniques to better understand the density and spatial distribution of news events. Heatmaps of story coverage, animations of incident locations over time, identification of most utilized places, and calendar indications of high incident days are all interesting analyses that could provide more insights to certain users (especially news editors, city officials, and those designing city interventions).

Perhaps the monitoring platform can eventually leverage other interesting information, such as demographic, crime, or weather data as potential additional features. However, other interesting analytical methodologies should leverage the data extraction functionality of the tool (future integration of download or API connections) to layer this georeferenced media data into systems more appropriate to handle this kind of data manipulation and yield additional GEOINT insights.

It may also be interesting to, much like the authors of (Teitler et al., 2008), explore the clustering of multiple articles into a single story, defined by spatial, temporal, and thematic proximities. This would facilitate more sophisticate analysis of the data, and identify those stories that reverberate through multiple agencies or are chronicled over time.

6.2.2 Develop a method for placemaking

Of particular value may be the creation of new features based on the compilation of similar features defined in the records of the Apregoar custom gazetteer. As new features are saved, their titles and descriptions could be automatically analyzed to identify common toponyms and potentially establish mutable areas that represent unofficial places referenced by city users. This could help to elucidate how city users view and describe different areas in Lisbon.

6.2.3 Facilitate ingestion of external features

Especially as organizations become more digitally savvy and/or dedicated to open data policies, there Apregoar publisher tool could benefit from the ingestion of existing features from other sources to ease the user experience of the journalist and provide more consistent georeferencing. Future iterations of the product should include a means by which shapefiles could be uploaded and saved into the progressive gazetteers, or connections to existing feature services may be created. This could be helpful when

referencing planned changes for the city, the footprints of which may exist in the GeoDados portal, or from other organizations that may already have feature definitions in their own systems that they are willing to share with the journalist. This feature is expected to support ease of use of the journalist, while also promoting confidence in the accuracy and precision of the geospatial annotations.

6.2.4 Deploy Apregoar tools

The web app should be publicly deployed for interaction with and contribution from any users. The Apregoar products should be licensed as free and open source such that it can be accessible and leveraged by other individuals and organizations for further development or related projects. Thus far, all sources of data are already publicly available, which mitigates concerns of distribution of proprietary materials. Future developments should continue to leverage open source tools, platforms, and data to support this end.

A plugin to relevant back offices (such as NewsPack on WordPress (Queiroz de Andrade & Carvalho, 2020)) should be developed and tested for easier integration of existing publishing processes. This should be tested with a local news provider to ensure viability in the market.

6.2.5 Distribute Apregoar data products

As was previously established, the Apregoar data products may contribute to the aggregation of public, private, and citizens knowledge such that it may be used for more nuanced applications in an increasingly spatialized context (Afzalan et al., 2017). One such application could be as a data layer to a local SDI to support "geomatics for sustainable societies" (Bhattacharya & Painho, 2018). Therefore, it is important that results are accessible by other users and applications (Shneiderman, 1996).

Users of any kind should be able to download files in comma separated value (CSV), GeoJSON, or shapefile formats. Additionally, results should be published via GeoRSS, such that they are easily searched, linked, and digested by other projects (Xing et al., 2015). An API should also be established so that other programs may make specific requests to extract data valuable to their applications.

6.2.6 Activate hyperlocality

Users may want to know what is happening near them. By introducing a feature that captures their current location and applies a radius, Apregoar could highlight articles that are spatially and temporally relevant in real-time.

Moreover, functionality could be developed such that passive location data is collected, and any incident occurring near to the user at the time of passing is also highlighted, creating a trail of relevant news stories for the user to review at their leisure.

6.2.7 Incorporate historical news

There may also be value in incorporating already published news stories into the geo-corpora, both for more complete and consistent coverage of news events over time and space at the more general level, but also for individual agencies who may wish to incorporate a more complete spatial and temporal repository of all of their digital media.

For each agency, if the yet un-geo-annotated corpora is small enough, it may be worthwhile to undergo a manual process of post-publication geo-annotation. However, for more general purposes or larger corpora, it may behoove the implementation of automated geo-annotation tools, such as those described in Section 2.3.3, (Halterman, 2019). The Apregoar toolset should create a means for integrating such auto-extracted georeferences (and mark them as such).

6.2.8 Improve user experience

Future iterations of the project should include an improved user experience via a more aesthetically pleasing and intuitive GUI. Data tells a story, and the Apregoar data products should help its users visualize this so that they may use it to apply further meaning to their own ends (Lupi, 2017).

In addition to the representation of news incidents as their polygon footprints, it may be appropriate to incorporate clustered representations of events when exploring larger areas or when many incidents are returned. In this case, the clustering methods could be incorporate the size of the footprints relative to the map object, such that larger polygons are preserved while the clustering permits easier zoom to smaller polygon representations.

6.2.9 Implement white-labeled experiences

In future variation, publishers should be able to 'white-label' their entries, creating a platform that they can integrate into their own websites that access and display only their own stories, as well as additional features such as recommending spatially or temporally similar articles to readers. These features, much like the now-common features of recommended stories by theme already integrated into may digital news sites, may support additional readership engagement with the site.

6.2.10 Recommend place

In the future, including a method that helps journalists to quickly identify potential spatial matches and ultimately confirm the suggestions or reject and design their own may also be a useful feature, much like in the some of the manual geo-annotation projects already being developed (Section 2.3.4). Based both on the techniques employed in these and the geographic scope of the agency (determined by their previous

geo-annotations and custom gazetteer entries), a recommendation system could be devised to suggest relevant locations to the journalist to improve the annotation process.

6.2.11 Expand language options

The web app should support the definition of use in English and Portuguese (leveraging a platform for expansion to other languages via internationalization and localization techniques) for all elements of the GUIs, such as project description, instructions, filters, units, etc. All data incorporated from external sources (such as news article contents, publisher tags, gazetteer names, etc.) may remain in their original forms/languages. If possible, alternate forms will be supported if provisioned by the original source. The language options of English and Portuguese should support the international use and cross investigation of a wider user base.

Bibliographic References

A Mensagem. (2022). Mensagem de lisboa - o novo jornal digital da cidade. Retrieved 2022-08-04, from https://amensagem.pt/. (Cit. on p. 25)

Acedo, A., Oliveira, T., Naranjo-Zolotov, M., & Painho, M. (2019). Place and city: Toward a geography of engagement. Heliyon, 5, e02261. https://doi.org/10.1016/j.heliyon.2019.e02261 (cit. on pp. xiv, 6, 7, 71, 72)

Afzalan, N., Sanchez, T. W., & Evans-Cowley, J. (2017). Creating smarter cities: Considerations for selecting online participatory tools. Cities, 67, 21–30. https://doi.org/10.1016/j.cities.2017.04.002 (cit. on pp. 8, 34, 35, 58, 70–72)

Al-Olimat, H. S., Thirunarayan, K., Shalin, V., & Sheth, A. (2018). Location name extraction from targeted text streams using gazetteer-based statistical language models, 1986–1997. http://arxiv.org/abs/1708.03105 (cit. on pp. ix, x, xiii, 6, 14, 73, 75)

AWS. (2021, October 20). Three-tier architecture overview. Retrieved 2022-11-04, from https://docs.aws.amazon.com/whitepapers/latest/serverless-multi-tier-architectures-api-gateway-lambda/three-tier-architecture-overview.html. (Cit. on p. 30)

Baker, J. M., Huddleston, G., & Atwood, E. (2019). The map as object: Working beyond bounded realities and mapping for social change. Educational Research for Social Change, 8, 138–152. https://doi.org/10.17159/2221-4070/2018/v8i1a9 (cit. on pp. 6, 7, 10, 11, 21, 53, 71, 72)

Barns, S. Joining the dots: Platform inermediation and the recombinatory governance of uber's ecosystem (M. Hodson, J. Kasmire, A. McMeekin, J. G. Stehlin, & K. Ward, Eds.). In: ed. by Hodson, M., Kasmire, J., McMeekin, A., Stehlin, J. G., & Ward, K. Routledge, 2020 (cit. on pp. 6, 69).

Bhattacharya, D., & Painho, M. (2018). Location intelligence for augmented smart cities integrating sensor web and spatial data infrastructure (smacisens). GISTAM 2018 - Proceedings of the 4th International Conference on Geographical Information Systems Theory, Applications and Management, 2018-March, 282–

289. https://doi.org/10.5220/0006786102820289 (cit. on pp. 1, 5–8, 10, 11, 31, 32, 53, 58)

Bonixe, L. (2019). As primeiras experiências de radiofusão local em portugal (1977-1984). Os média no Portugal Contemporâneo, 19, 183–195. https://doi.org/https://doi.org/10.14195/2183-5462_35_12 (cit. on p. 23)

Brown, G. G., & Pullar, D. V. (2012). An evaluation of the use of points versus polygons in public participation geographic information systems using quasi-experimental design and monte carlo simulation. International Journal of Geographical Information Science, 26, 231–246. https://doi.org/10.1080/13658816.2011.585139 (cit. on pp. ix, xiv, 6, 7, 21, 52, 72)

Börner, K., Bueckle, A., & Ginda, M. (2019). Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. PNAS, 116, 1857–1864. https://doi.org/10.1073/pnas.1807180116 (cit. on pp. 9, 10)

Cai, G., & Tian, Y. (2016). Towards geo-referencing infrastructure for local news. Proceedings of the 10th Workshop on Geographic Information Retrieval, GIR 2016, 1–10. https://doi.org/10.1145/3003464.3003473 (cit. on pp. 5, 11–15, 17, 73, 75, 79)

Chiappinelli, C. (2020, November). Democracy: Fueled by pizza. Retrieved 2020-12-17, from https://www.esri.com/about/newsroom/publications/wherenext/pizza-to-the-polls-on-election-day/. (Cit. on p. xiv)

Creative Commons. (n.d.). What we do - create commons. Retrieved 2022-08-09, from https://creativecommons.org/about/. (Cit. on p. ix)

Câmara Municipal de Lisboa. (n.d.). Notícias. Retrieved 2022-08-04, from https://www.lisboa.pt/atualidade/noticias. (Cit. on p. 24)

Datta, A. (2018). Top six geoint trends. Retrieved 2020-12-18, from https://www.geospatialworld.net/blogs/top-six-geoint-trends/. (Cit. on pp. xi, 10, 11, 53)

DGT. (2022). Carta administrativa oficial de portugal. Retrieved 2022-08-04, from https://www.dgterritorio.gov.pt/cartografia/cartografia-tematica/caop?language=en. (Cit. on p. 25)

DMA WGS 84 Development Committee. (1991, September). Department of defense world geodetic system 1984. The Defense Mapping Agency. https://apps.dtic.mil/sti/pdfs/ADA280358.pdf. (Cit. on pp. xvi, 5)

ECMA International. (2017). Json. Retrieved 2022-08-23, from https://www.json.org/json-en.html. (Cit. on p. xii)

Eisl, M. (2020). Searching european data. (Cit. on p. 69).

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. Annals of the American Association of Geographers, 1–20. https://doi.org/0.1080/00045608.2011.595657 (cit. on pp. xiii, xvi, 13, 14, 71)

ESRI. (n.d.). Adjust how locations and attributes are extracted. Retrieved 2021-08-03, from https://pro.arcgis.com/en/pro-app/latest/help/data/locatext/adjust-how-locations-and-attributes-are-extracted.htm. (Cit. on p. 78)

European Commission. (n.d.). Background - nuts - nomenclature of territorial units for statistics - eurostat. Retrieved 2022-08-08, from https://ec.europa.eu/eurostat/web/nuts/background. (Cit. on p. xiii)

Eurostat. (n.d.). Eurostate regions and cities - overview. Retrieved 2021-08-25, from https://ec.europa.eu/eurostat/web/regions-and-cities/overview. (Cit. on p. 21)

Evans-Cowley, J., & Hollander, J. (2010). The new generation of public participation: Internet-based participation tools. Planning Practice and Research, 25, 397–408. https://doi.org/10.1080/02697459.2010.503432 (cit. on pp. 9, 52, 71, 72)

Fitoussi, E. (n.d.). Geo my wordpress. Retrieved 2020-12-18, from https://wordpress.org/plugins/geo-my-wp/. (Cit. on p. 79)

Free Software Foundation. (2022, June 25). What is free software? Retrieved 2022-10-31, from https://www.gnu.org/philosophy/free-sw.en.html. (Cit. on p. x)

Freguesia de Estrela. (2022). Freguesia de estrela. Retrieved 2022-08-04, from https://www.jf-estrela.pt/. (Cit. on p. 25)

Giorgia lupi. (n.d.). Retrieved 2021-01-04, from https://www.stories.com/en_eur/giorgialupi.html. (Cit. on p. 9)

Google News Initiative. (2018). Elevating quality journalism digital news innovation fund report 2018. Digital News Innovation Fund. (Cit. on pp. 9, 11, 12, 23).

Granger, J. (2020). Isabelle roughol of borderline podcast, on the pros and cons of 'indie journalism'. Journalism.co.uk. https://www.journalism.co.uk/podcast/isabelle-roughol-of-borderline-podcast-on-the-pros-and-cons-of-indie-journalism-/s399/a766817/ (cit. on p. 53)

Gritta, M, Pilehvar, M., & Collier, N. (2018). 1, 1285–1296. https://doi.org/10.18653/v1/p18-1119 (cit. on pp. 13, 14, 73, 75)

Gupta, S., & Nishu, K. Mapping local news coverage : Precise location extraction in textual news content using fine-tuned bert based language model. In: Association for Computational Linguistics, 2020, 155–162 (cit. on pp. xi, 12–14, 73, 75).

Halterman, A. Geolocating political events in text. In: Association for Computational Linguistics, 2019, 29–39. https://doi.org/10.18653/v1/w19-2104 (cit. on pp. xv, 13–16, 59, 76).

Hamborg, F., Breitinger, C., & Gipp, B. Giveme5w1h: A universal system for extracting main events from news articles. In: INRA, 2019 (cit. on pp. 12–15, 78).

Hintz, D., & Hantke, C. (2020). How government agencies are integrating & delivering data for emergency response. https://www.gotostage.com/channel/0e36a00d2d1942a094e92ef2536154ed/recording/9895203fea574a38868d14ca66d41bd9/watch. (Cit. on p. 8)

IBM Cloud Education. (2020, October 28). Three-tier architecture. Retrieved 2022-11-04, from https://www.ibm.com/cloud/learn/three-tier-architecture. (Cit. on p. 30)

Idealista. (n.d.). Mapa de campolide, lisboa: Casas à venda. Retrieved 2022-08-03, from https://www.idealista.pt/comprar-casas/lisboa/campolide/mapa. (Cit. on p. 24)

Imani, M. B., Khan, L., & Thuraisingham, B. (2019). Where did the political news event happen? primary focus location extraction in different languages. Proceedings - 2019 IEEE 5th International Conference on Collaboration and Internet Computing, CIC 2019, 61–70. https://doi.org/10.1109/CIC48465.2019.00017 (cit. on pp. 7, 12, 13, 15, 16, 73, 75, 77)

Institudo Nacional de Estatística. (2021). Censos 2021 resultados prelíminares. Institudo Nacional de Estatística. Retrieved 2022-08-02, from https://www.ine.pt/scripts/db_censos_2021.html. (Cit. on pp. 21, 25)

James. (2020). An attempt to extract geo-location from text. Retrieved 2020-12-18, from https://medium.com/datadriveninvestor/an-attempt-to-extract-geo-location-from-text-c76cb6bd49d4. (Cit. on pp. 14, 73, 75)

Jiang, H., Genderen, J. V., Mazzetti, P., Koo, H., Chen, M., Jiang, H., Genderen, J. V., Mazzetti, P., Koo, H., Chen, M., & Koo, H. (2020). Current status and future directions of geoportals. International Journal of Digital Earth, 13, 1093–1114. https://doi.org/10.1080/17538947.2019.1603331 (cit. on pp. xi, xv, 6–8, 10, 18, 32, 34, 53, 70)

Junta de Freguesia de Campo de Ourique. (2022). Junta de freguesia de campo de ourique. Retrieved 2022-08-04, from https://www.jf-campodeourique.pt/. (Cit. on p. 25)

Junta de Freguesia de Campolide. (2022). Junta de freguesia de campolide. Retrieved 2022-08-04, from https://www.facebook.com/jfcampolide. (Cit. on p. 25)

Karimzadeh, M., & MacEachren, A. M. (2019). Geoannotator: A collaborative semi-automatic platform for constructing geo-annotated text corpora. ISPRS International Journal of Geo-Information, 8. https://doi.org/10.3390/ijgi8040161 (cit. on pp. xi, 13, 16, 30, 79)

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrun, J. O. (2019). Geotxt: A scalable geoparsing system for unstructured text geolocation. GIS, 23. https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12510 (cit. on pp. x, 12–14, 16, 73, 75, 76)

Kleinhans, R., Ham, M. V., & Evans-Cowley, J. (2015). Using social media and mobile technologies to foster engagement and self-organization in participatory urban planning and neighbourhood governance. Planning Practice and Research, 30, 237–247. https://doi.org/10.1080/02697459.2015.1051320 (cit. on pp. 6, 8, 72)

Lee, S. J., Liu, H., & Ward, M. D. (2019). Lost in space: Geolocation in event data. Political Science Research and Methods, 7, 871–888. https://doi.org/10.1017/psrm.2018.23 (cit. on pp. x, xii, 2, 13–16, 76, 78, 83)

Leszczynski, A. (2019). Platform affects of geolocation. Geoforum, 107, 207–215. https://doi.org/10.1016/j.geoforum.2019.05.011 (cit. on pp. xii, 7, 12, 69)

Lieberman, M. D., Samet, H., & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. Proceedings - International Conference on Data Engineering, 201–212. https://doi.org/10.1109/ICDE.2010.5447903 (cit. on pp. xii, xv, 1, 13–17, 21, 73)

Longley, P., Goodchild, M., D., M., & Rhind, D. (2005). Geographic information systems and science (2nd ed.). John Wiley & Sons, Ltd. http://www.biouls.cl/~david/sig/gisandscience2ed.pdf. (Cit. on pp. 5, 7)

Lourenço, J. M. (2021). The NOVAthesis LaTeX Template User's Manual. NOVA University Lisbon. https://github.com/joaomlourenco/novathesis/raw/master/template.pdf. (Cit. on pp. ii, iv)

Lupi, G. (2017). Data humanism, the revolutionary future of data visualization. Retrieved 2021-01-04, from https://www.printmag.com/post/data-humanism-future-of-data-visualization. (Cit. on pp. 9, 10, 59, 70)

Marshall, S. (2012). #Tip of the day for journalists: Geotag your content using a wordpress plugin. Journalism.co.uk. https://www.journalism.co.uk/tip-of-the-day/-tip-of-the-day-for-journalists-geotag-your-content-using-a-wordpress-plugin/s419/a561891/ (cit. on pp. 12, 17)

Massey, D. (1991). A global sense of place. Marxism Today, June, 24–29. http://banmarchive.org.uk/collections/mt/pdf/91_06_24.pdf (cit. on pp. 5, 7–9)

Meeks, E. (2019, December). 2019 was the year data visualization hit the mainstream. Retrieved 2020-09-25, from https://medium.com/nightingale/2019-was-the-year-data-visualization-hit-the-mainstream-d97685856ec. (Cit. on pp. 9, 10, 34)

Monmonier, M. (2018). How to lie with maps (3rd ed.). The University of Chicago Press. (Cit. on p. 10).

Nordquist, R. (2018). Exonym and endonym. Retrieved 2020-12-18, from https://www.thoughtco.com/exonym-and-endonym-names-1690691. (Cit. on p. x)

Oliveira, T. H. M. D., & Painho, M. Open geospatial data contribution towards sentiment analysis within the human dimension of smart cities (A Mobasheri, Ed.). In: ed. by Mobasheri, A. Springer International Publishing, 2021, pp. 75–95. isbn: 9783030582326. https://doi.org/10.1007/978-3-030-58232-6 (cit. on pp. 9, 31, 70).

Open Knowledge Foundation. (n.d.-a). Open data commons open database license (odbl). Retrieved 2022-08-09, from https://opendatacommons.org/licenses/odbl/. (Cit. on p. xiv)

Open Knowledge Foundation. (n.d.-b). Open definition. Retrieved 2022-10-21, from https://opendefinition.org/. (Cit. on pp. xiii, 68)

OpenStreetMap contributors. (2022). Openstreetmap. Retrieved 2022-08-04, from https://www.openstreetmap.org/. (Cit. on p. 27)

Painho, M., & Pina, I. (2013). The invisible cities-can ppgis connect citizens to urban policies? Revista Internacional de Ciencia y Tecnología de la Información Geográfica, 1, 1–4 (cit. on pp. 7, 8, 11, 52, 72).

Pallets. (n.d.-a). Flask | the pallets projects. Retrieved 2022-08-10, from https://palletsprojects.com/p/flask/. (Cit. on p. 31)

Pallets. (n.d.-b). Jinja | the pallets projects. Retrieved 2022-08-10, from https://palletsprojects.com/p/jinja/. (Cit. on p. xii)

Pallets. (n.d.-c). Werkzeug | the pallets projects. Retrieved 2022-08-10, from https://palletsprojects.com/p/werkzeug/. (Cit. on pp. xvi, 32)

Panchaud, N. H., & Hurni, L. (2018). Integrating cartographic knowledge within a geoportal: Interactions and feedback in the user interface. Cartographic Perspectives. https://doi.org/10.14714/CP89.1402 (cit. on p. 30)

PostgreSQL Tutorial. (2022). Postgresql tutorial. Retrieved 2022-08-03, from https://www.postgresqltutorial.com/. (Cit. on pp. xv, xvi)

QGIS Development Team. (n.d.). Qgis (Version 3.18.3). https://download.qgis.org/downloads/. (Cit. on p. xiv)

QGIS project. (2022, May 18). A gentle introduction to gis (3.22). Retrieved 2022-10-14, from https://docs.qgis.org/3.22/en/docs/gentle_gis_introduction/. (Cit. on p. ix)

Queiroz de Andrade, D., & Carvalho, C. (2020). Interview. (Cit. on pp. 58, 82).

Rajabifard, A. (2009). Realizing spatially enabled societies – a global perspective in response to millennium development goals. Eighteenth United Nations regional Cartographic, Conference for Asia and the Pacific, Bangkok, 26-29 October, 9. https://unstats.un.org/unsd/geoinfo/RCC/docs/rccap18/IP/18th_UNRCCAP_econf.100_IP4.pdf (cit. on pp. 6, 71)

Rivera, G., Florencia, R., García, V., Ruiz, A., & Sánchez-Solís, J. P. (2020). News classification for identifying traffic incident points in a spanish-speaking country: A real-world case study of class imbalance learning. Applied Sciences (Switzerland), 10. https://doi.org/10.3390/APP10186253 (cit. on pp. 9, 36, 71)

Roche, S., & Rajabifard, A. (2012). Sensing places' life to make city smarter. ACM SIGKDD Interntional Workdshop on Urban Computing (UrbComp 2012) (cit. on pp. xv, 5–8, 12, 70–72).

Sami, R. (2019, October). Tools i recommend for building geospatial web applications | by ramiz sami | the startup | medium. Retrieved 2020-09-25, from https://medium.com/swlh/tools-i-recommend-for-building-geospatial-web-applications-274d6939536c. (Cit. on pp. 30–32)

Shneiderman, B. (1996, July). The eyes have it: A task by data type taxonomy for information visualizations. University of Maryland. http://www/cs.umd.edu/projects/hcil/. (Cit. on pp. 9, 10, 58)

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., & Cardoso, N. (2006). Adding geographic scopes to web resources. Computers, Environment and Urban Systems,

30, 378–399. https://doi.org/10.1016/j.compenvurbsys.2005.08.003 (cit. on pp. xiii–xv, 2, 13, 73, 74, 78)

Snyder, L. S., Karimzadeh, M., Chen, R., & Ebert, D. S. (2019). City-level geolocation of tweets for real-time visual analytics. arXiv, 0–3 (cit. on pp. 10, 15, 20).

Stamen Design. (n.d.). Stamen toner lite. Retrieved 2022-08-04, from http://maps.stamen.com/toner-lite/#12/37.7706/-122.3782. (Cit. on p. 27)

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., & Sperling, J. (2008). Newsstand: A new view on news. GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, 144–153. https://doi.org/10.1145/1463434.1463458 (cit. on pp. xiii, 11–13, 17, 31, 52, 57, 73, 77)

Tobler, W. (1970). A computer movie simulating urban growth in the detroit region. Economic Geography, Supplement: Proceedings. International Geographic Union. Commission on Quantitative Methods, 46, 235–240. http://www.jstor.org/stable/143141 (cit. on p. xv)

Varanda, A. J. A. (2020). Project "oraculo": Extracting events from news streams and mining their spatiotemporal patterns to support un operations in the central african republic. https://www.gotostage.com/channel/0e36a00d2d1942a094e92ef2536154ed/recording/7f595876d4e84d618db34217386854f3/watch?source=CHANNEL. (Cit. on pp. 10, 11, 53)

Williams, P. (2016). What, exactly, is a smart city? Retrieved 2020-11-28, from http://meetingoftheminds.org/exactly-smartcity-16098. (Cit. on p. 70)

Witschas, S. Cross-border mapping-geodata and geonames. In: Borders in a new Europe, 2004 (cit. on pp. x, 7, 13, 73).

World Economic Forum, & ScaleUpNation. (2021). Circular trailblazers: Scale-ups leading the way towards a more circular economy, World Economic Forum. (Cit. on pp. 9, 69).

Xing, H., Chen, J., & Zhou, X. (2015). A geoweb-based tagging system for borderlands data acquisition. ISPRS International Journal of Geo-Information, 4, 1530–1548. https://doi.org/10.3390/ijgi4031530 (cit. on pp. xi, xiii, 6–8, 11, 12, 58)

Zhang, Y., Chanana, K., & Dunne, C. (2019). Idmvis: Temporal event sequence visualization for type 1 diabetes treatment decision support. IEEE Transactions on Visualization and Computer Graphics, 25, 512–522. https://doi.org/10.1109/TVCG.2018.2865076 (cit. on p. 10)

A

Additional context on geospatial information

A.1 Open geospatial standards

The OGC is a global endeavor that joins businesses, government agencies, and universities across the globe. Their efforts develop relevant standards for spatial information such that it is "findable, accessible, interoperable, and reusable" (Open Knowledge Foundation, n.d.-b). Similar principals underpin the concept of open data, in which data is provided without restrictions of access or use. Though OGC boasts has developed many standards, APIs, best practices, papers, etc., some of the most relevant to this project are WMS and WFS, both of which are accessed via HTTP interfaces.

WMS provides map images via the "GetMap" operation that can be displayed in a web browser. Users may specify various parameters such as the size and styling, SRS, bounding boxes, etc. Providers can configure the settings and offer of data access, which a user can view via the "GetCapabilities" operation, which returns an XML file. By requesting "GetFeatureInfo", users can access additional attributes of the feature(s) found at a particular coordinate (Open Knowledge Foundation, n.d.-b).

WFS provides a means for users to access and even manipulate the served geographic information via a web browser. Depending on how the service is configured (accessible via the "GetCapabilities" request), users may be able to access some or all of the CRUD functions. Via the "GetFeature" request, users can access Geography Markup Language (GML) or GeoJSON formatted contents of the requested features, including geocoordinates. Additional operations allow for various versions of discovery, query, stored query, locking, and transaction with the served data (Open Knowledge Foundation, n.d.-b).

These services can be leveraged by third party applications in conjunction with other datasets and operations for further processing or visualizations.

A.2 Commercial geospatial platforms

Commercial geospatial platforms leverage geospatial data as an ”economic commodity” (Leszczynski, 2019) to provide value-added services to their customers. In smart environments, geospatially enabled devices and systems monitor and orchestrate automatic responses to the environment without the need of user input (Leszczynski, 2019). Simultaneously, various online services and products use geospatial awareness to enhance interactive objects that may attract users and extend their visit duration, via spatially attuned experiences utilizing LBS (Leszczynski, 2019). Some of the most familiar business-to-consumer (B2C) applications include those that provide navigation (Google Maps), ride share (Uber), food ordering (Glovo), accommodation seeking (Idealista, AirBnB), and dating (Happn) services. Often for services, these systems will apply hyperlocality by physically and temporally center the experience to the user, while for networks, the inclusion of geolocation fosters a sense of connectedness with other users throughout the community or world (Leszczynski, 2019). Most of these incorporate some sort of real time detection of user location, usually via GPS, though it could also leverage various proximity sensors or remote sensing techniques. They might also or instead provide the opportunity to user-define areas of interest for spatial searching or filtering (Barns, 2020). This may be via a form, in which other attributes (TA and ThA) may be incorporated, or by selecting points or drawing polygons within which to focus the search (Eisl, 2020).

Digital platform ecosystems are an extension of geoplatforms in that they ”[interconnect] stakeholders and organizations exchanging knowledge and value, each with their own role in the greater whole” (World Economic Forum & ScaleUpNation, 2021). Effectively, these ecosystems aim to create a fluid user experience and add even more value to any individual product by incorporating relevant information from other incorporated systems (think the alternate methods of navigating in Google Maps, which uses data from other businesses to present transit times and may include real time costs of ridehailing services or soft mobility options). Though incredibly interesting, these are outside of the scope of the thesis.

B

Additional context on place and people

B.1 Smart communities

Though most often referred to on the city level, smart communities leverage ICT (networks) and various sources of data (sensors) to address and improve the functional needs of its population (actuators), engaging them to develop citizen-centered interventions and responding to their changing needs (Afzalan et al., 2017; Roche & Rajabifard, 2012; Williams, 2016). In fact, "[a]n active and engaged citizen is indeed the main driving force of a 'smart city'" (Oliveira & Painho, 2021). Though smart cities also address economic vitality and environmental impact in addition to social well-being, empowered communities increasingly expect the ability to influence their environments, such as by affecting government planning procedures and services (Williams, 2016). Beyond efficiency, citizens require safer, more enjoyable living experiences in all aspects of their lives. Governments may accommodate the public interest (Afzalan et al., 2017) by incorporating four dimensions (intelligence, digital, open, and live, referring to its social and informational infrastructures, open governance, and continuity of adaptation, respectively) of smart communities (Oliveira & Painho, 2021). The identification and monitoring of community dynamics requires the sensing of life though open dialogues with constituents as well as the employment of IOT technologies (Roche & Rajabifard, 2012). This sensing infrastructure leverages multiple sources of data to determine the state of various subsystems and support interventions. Ideally, it would identify potential opportunities for improvement but is more commonly leveraged in application focused scenarios, in which a "search, evaluate, and process" method is employed in response to a particular challenge (Jiang et al., 2020).

Data in general is already highly regarded as a key commodity for developing an economy (Lupi, 2017). To harness the value of this ever-expanding resource, community operations should accommodate methods for capturing, exploring, and sharing this data, spatial or otherwise, and its processed results (Roche & Rajabifard, 2012). Beyond operational efficiency, the information products and services have the potential to stimulate new creative uses that facilitate the economic, social, and environmental

well-being of the community (Rajabifard, 2009). Just as the context of a community – its culture, history, environment, access to technology, demographics, etc. – can vary tremendously across time and space, so too should its interventions (Afzalan et al., 2017). Members of such "knowledge societies" (Rivera et al., 2020), investigators and entrepreneurs or anyone with access to technology, are better equipped to address local psychographics in nontraditional or niche applications (Baker et al., 2019). Such opportunities can even unburden institutions with the responsibility of managing, processing, and transforming data into relevant services, and instead allow the community itself to develop novel applications for public resource that can be adapted into operations when mature.

B.2 Public participation

Public participation is a critical element of citizen empowerment, democratic vibrancy, and innovation (Afzalan et al., 2017). It provides opportunities for citizenry to provide feedback on services and provide new ideas based on lived realities, but also opportunities for collaboration and motivated co-productions with interested, non-institutional stakeholders within the area (Acedo et al., 2019). Further, participation strengthens a community by building social capital amongst its participants, demonstrating trust between members (Evans-Cowley & Hollander, 2010). High forms of civic engagement assume that citizens have the power to influence decisions that will touch their own lives, whether through active dialogues or other means of engagement. Though not a new concept, today's communities are more and more expecting that relevant organizations will provide opportunities for such feedback, which (if implemented appropriately) may harness public knowledge for the better of said organization and the community as a whole. Updated strategies, especially those that include in-person and digitally hybrid participation options, may engage larger audiences, facilitating greater participation while mitigating possible digital divides in participating demographics (Afzalan et al., 2017; Elwood et al., 2012; Evans-Cowley & Hollander, 2010). It can also prolong interactions, allowing all stakeholders to reevaluate options and motivations throughout the entire process (Afzalan et al., 2017). These services may be government or institutional services (ex: Lisboa Participativa), non-institutional offerings (ex: PlaceSpeak) or commercial products leveraged for engagement (ex: NextDoor).

As an "inherently spatial" element, public participation should not be disconnected from its geo-dimension (Acedo et al., 2019). Spatial information is critical for making educated decisions on key human issues (Rajabifard, 2009). Though this clearly applies to decisionmakers in their respective fields, access to location services is must be a given for all modern society (Rajabifard, 2009). "[A] city could not be smart without spatially enabled citizens" (Roche & Rajabifard, 2012), who are able to contextualize their own experiences and needs in relation to the realities of their peers. In 1996, this was recognized by the National Center for Geographic Information and Analysis

in the United States of America which established the PPGIS to better accommodate marginalized populations (Brown & Pullar, 2012). It can be especially powerful to visualize the impact of interventions of underrepresented communities at scale (Baker et al., 2019). At its core, a PPGIS represents an abstract of thematically interesting features, contributing to a communal understanding of place (Brown & Pullar, 2012). "[I]n spite of all the technological developments in recent years, one of the biggest barriers to public participation in urban policies remains unsurpassable: the difficulty that people have to understand how the planning proposals are projected in space, how they redefine it, and how they impact the use of urban space" (Painho & Pina, 2013). This kind of technology supplements top down and bottom up activism to provide a common foundation from which to build collaborative understanding and develop effective interventions through "active citizenship" (Kleinhans et al., 2015). From this, such online tools should include elements of understanding the decision making processes and tracking its progress. Both of these support transparency and opportunities to influence decisions via connection, sharing of information, and a platform for developing ideas (Afzalan et al., 2017).

A critical element of any spatial understanding, smart or participative, is the collection of data with a geospatial element. Beyond intermittent and representative polls of the community and implanted IOT devices capturing objective states of the environment, citizens themselves are a wealth of spatially routed information within a community (Roche & Rajabifard, 2012). Whether actively shared or passively shared data, and whether primarily focused towards citizen engagement (such as answering a poll) or extracted for such use (such as sentiment extraction of public social media posts), VGI is critical to the understanding of "citizens' social synergies in the urban context" (Acedo et al., 2019; Evans-Cowley & Hollander, 2010). Especially in issues of public planning, the understanding of individuals' spatial context can realign the lens, and therefore the results, of community initiatives towards the people of whom it is composed. Though there continues to be a disparity between the understanding of places and the people who inhabit them (Acedo et al., 2019), these tools can establish a better connection between the where and the why and how of spatial phenomenon and the perspectives of those who experience them (Painho & Pina, 2013).

C

Additional context on geoparsing

C.1 Toponym disambiguation

A prominent challenge of geoparsing is disambiguation of toponyms extracted from unstructured text data from potential matches in the accessed gazetteer(s). Not only may one toponym reference multiple places or entities of various types (referent ambiguity or referent class ambiguity, respectively), but one place may be known by multiple names (reference ambiguity) (Al-Olimat et al., 2018; Cai & Tian, 2016; Gritta et al., 2018; Gupta & Nishu, 2020; Imani et al., 2019; James, 2020; Karimzadeh et al., 2019; Lieberman et al., 2010; Silva et al., 2006; Teitler et al., 2008; Witschas, 2004). This is further confounded by the employment of nameheads, which may result in appellation formation (such as replacing THE 25 OF APRIL BRIDGE with THE BRIDGE), explicit metonymy (UNIVERSITY OF NOVA vs. NOVA), which require additional context to resolve (Al-Olimat et al., 2018). Natural language frequently employs shortened versions of names in contexts where the implied words can be dropped. Social media provides both an additional challenge of character limitations (either platform imposed or as a common practice, which introduces unconventional phrase shortening tactics), as well as the potential for additional context (such as leveraging their profile information or post geotags to provide additional spatial queues).

These nameheads are also associated with delimitation challenges, such as category ellipsis (CITY OF LISBON vs. LISBON), and/or location ellipsis (ACADEMY OF SCIENCES OF LISBON vs. ACADEMY OF SCIENCES), which may be address via statistical language models (Al-Olimat et al., 2018). However, the challenge of delimitation may extend beyond the application of potential toponym matches' boundaries. Especially at small scales, there may not be an existing, static boundary in an external gazetteer that matches what the author intends to describe. This could either be due to a high level of granularity in the text descriptions (such as a particular intersection) (Gupta & Nishu, 2020), or because the specific area of interest to the author does not have an official location association (neighborhoods that exist culturally but are not defined at any administrative level). These may be considered wandering, as they

may not be consistent across understandings or may morph over time according to the common vernacular (Silva et al., 2006).

For example: Lisbon, Portugal has several understandings associated with it. There are officially delineated areas, such as Lisbon municipality, AML, and the District of Lisbon. There are also unofficial associations defined by functional city topography, such as the land contained between the Tejo River and either the 2nd Circular or Circular Regional Interior de Lisboa (CRIL) (Figure C.1). Users of the city (inhabitants, workers, enjoyers, etc.) may refer to any of these areas (or others) as "Lisbon", either because this simplification is clear in context, or because they have confused the actual city boundaries. Note that neither area includes all of the actual City of Lisbon area – CRIL bisects the freguesia of Parque das Nacões but also includes parts of neighboring municipalities (Amadora, Odivelas, and Loures), and the 2nd Circular road cuts out some of the more distal areas of the city (the freguesias of Parque das Nações, Olivais, Santa Clara, Lumiar, Carnide, as well as part of Benfica).

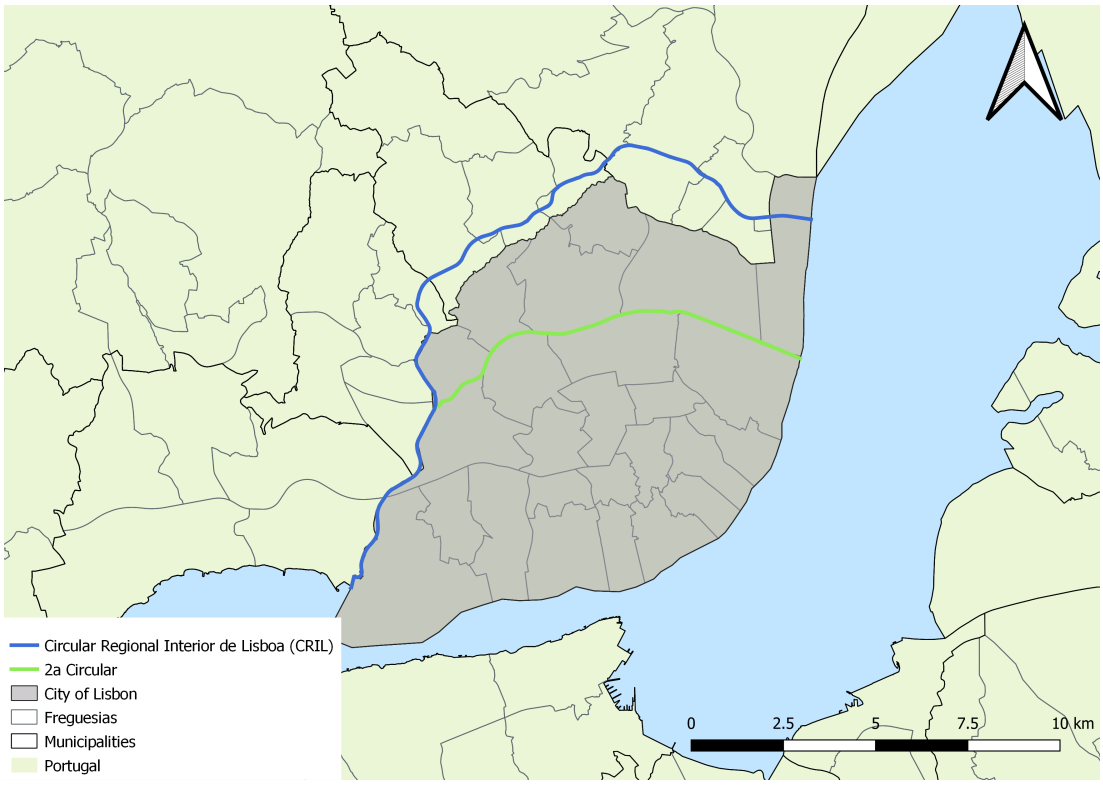

Figure C.1: Multiple colloquial definitions of "Lisbon"

C.2 Geoparsing tools

The challenge of developing tools by which to automate the geoparsing of natural language texts is longstanding. Efforts in identifying place name (Standford Core NLP,

Apache OpenNLP, Cliff-Clavin, Mordecai) and further to disambiguate between potential matches (CLAVIN, Geoparser.io, Geography3, Edinburgh Geoparser, GeoTxT, CamCoder) are continually refining the extraction of place by the application of various combinations of heuristics, statistics, and machine learning strategies with ever increasing resources (Al-Olimat et al., 2018; Gritta et al., 2018; Imani et al., 2019; James, 2020; Karimzadeh et al., 2019). However, as they "reflect human conceptualization and experiences of space and places... [these] text-based spatial descriptions are subject to all sorts of ambiguities that prevent effective use" (Cai & Tian, 2016). Some of these can be overcome by finetuning on target datasets (Gupta & Nishu, 2020), while others cannot yet be reliably overcome with current resources on a large scale. Additional challenges include definitions of hierarchy, fuzzy conceptualizations, and scale.

D

Examples of geo-annotation in literature

D.1 Automatic geo-annotation examples

Many efforts, both commercial and academic, have been undertaken to extract event (or sometimes only geographic) information from digital media texts. Many of these rely on sophisticated NLP techniques for pattern recognition and classifiers, local lexicons, event resolution and de-duplication, and multi-language support to be useful in automatic event extraction (Halterman, 2019; Lee et al., 2019). Examples follow.

The GDELT Project is a project that extracts place as well as actors, sentiment, and event connection (among other elements) from journalistic media across the globe, including publications from as far back as 1979. This and similar projects are powerful and hugely informative, especially as they apply to existing published data. However, the existing automated extraction includes several challenges: i) it often mis-attributes identified toponyms , ii) it does not support the subtlety of incidents occurring in non-conforming places (an incident may not apply to a single administrative boundary but really fall into a subsection of one or several); iii) it requires technical prowess and tools to explore the data. A user is unable to define a spatial area of interest (such as their route to work with a half mile buffer zone or some other irregular shape) and search for all spatially related results, nor it is easy to apply TAs or ThAs without prior experience querying results. It also fails to link documents to their geolocation by using additional semantic tools, rather returning simply the main location from the reviewed sentence (Halterman, 2019). The project is a valuable tool for summarizing global scale events/tendencies with an output format suitable for more formal investigations and experienced geographers.

GeoTxt (Karimzadeh et al., 2019), a geoparser for unstructured text, allows users to select from one of six NER engines as well as filter for spatial local or attribute parameters. Optionally, a user can apply additional co-occurring, hierarchical, or spatial proximity-based disambiguation mechanisms. The resulting ranked results from the GeoNames gazetteer should better disambiguate recognized toponyms and allow less arduous association of geographic footprints to documents. The program was tested

on global tweets, and considered each as a separate document (unrelated to others written by the same user). Its most promising results used the CogComp or Stanford NERs, though the authors note that tweets underperform in comparison to long text, and recommend either building a tweet specific training corpus or improvement of the preprocessing pipelines. They also found that heuristics prioritizing highly populated toponyms and feature codes improved results, while hierarchical and proximity-based methods decreased accuracy.

Profile, developed by (Imani et al., 2019), is a geoparser that extracts focus locations from political news reports in multiple languages (English, Spanish, and Arabic). To do this, they leverage NER to identify candidate locations and the sentences in which they appear, from which they extract semantic features, and then finally predict the primary focus area using a trained classifier on prelabeled instances of different languages. Their results outperformed the existing methods because of the use of semantic relationships, with best performances seen in English. Training and testing, however, was limited to certain kinds of news topics, assumed a single event per article, and requires training on the same dataset that will be tested (a training dataset from The New York Times would not translate well in tests of The Huffington Post, for example).

NewsStand is a spatio-textual news aggregator and display interface created by (Teitler et al., 2008) that retrieves, analyzes, and maps news stories, allowing users to peruse data by topic or location, with different renderings of news stories based on zoom level and pan position while balancing story significance. The application was created to address common spatial queries: Where did this happen? and What happened here? The main objective is to create a UI that can convey as much information as possible, always maintaining a full window of various stories regardless of where one is navigating. When the visible map area represents too many stories to be comfortably ingested at one time, it prioritizes the more globally interesting (”significant”) documents and limits these to a maximum number of markers per subarea of the screen. The interactive interface allows users to identify stories by location and open them to uncover more specific contextual maps. Users may zoom to globe, country, state, or city levels. Underlying this functionality is a data collection process (ingestion of RSS feeds), then each article is geoparsed and its geographic focus determined. Unlike other systems evaluated here, however, NewsStand does not map the locations of articles, but rather story clusters (potentially multiple articles and/or sources discussing the same event). This is done by grouping articles with similar ”story content” and ”story lifetime” (proximate publication dates). Then, the geographic focus of the story cluster is computed. The display uses point representations corresponding to the GeoNames gazetteer matches for geographic focuses, which is appropriate for an application leveraging article clustering, however it may obscure the actual areas associated with each event.

LocateXT is a software extension for the ArcGIS suite of tools. It advertises that it searches unstructured data for spatial locations, then generates point features (no

line or polygon options are available) representing those locations. In addition to the ArcGIS gazetteers, users may supply their own custom locations gazetteers to augment the out-of-the-box features. The system does not automatically geocode addresses (a different tool in their suite is available for this). If recognized, recent dates may be associated as TAs. In unstructured text data, other thematic data (keywords) may be associated as additional attributes. For semistructured data, custom attributes and rules to fill them may be defined (ESRI, n.d.).

Tumba, a web search engine for Portugal, sought to improve its functionality by including geographical knowledge inferred from web sources (Silva et al., 2006), In 2006, the researchers noted an average of 2.2 geographic references per online document related to any of the Portuguese municipalities, demonstrating the "pervasive" nature of geographic information on the web. They attempted to retroactively define a geographic scope of existing online articles, and match these with the footprints of toponyms used in search queries. The system is no longer active.

Giveme5W1H, an open source tool developed by (Hamborg et al., 2019), seeks to extract all key defining elements of an event: who, what, where, when, how, why (5W1H). The system ingests an article, applying NLP and normalizing the results. Dates are parsed (including relative to publish dates), and resolved to a single point in time. Locations are geoparsed, actors are identified, and phrases are extracted to retrieve phrases answering 'how?'. Finally, the candidate answers of each 5W1H question are scored based on described lingual heuristics.

(Lee et al., 2019) attempt to innovate on existing models by applying custom gazetteers to news about areas in conflict. Unlike other tools that assume a singular geographic event focus, their works seek to identify all focus localities. To do this, they built custom dictionaries: locations, actors, relevant protest words; and relevant fight words. These dictionaries are then applied to the text and N-grams are used to reveals patterns that can be used to identify event-relevant information. Their approach improved classification of results by 25% over traditional dictionary approaches.

Beyond academic studies, several commercial endeavors are leveraging geolocation extraction to provide new services to readers. InYourArea is digital news aggregator that seeks to serve spatially relevant news to their users (the United Kingdom), by requiring at least one associated postal code (though more may be associated if preferred). It is unclear how they geotag their news articles, though because of the narrow scope, single language, and presumably limited number of publishers, it is possible that they leverage a particularly robust local gazetteer. The UI spotlights a Newsfeed with several more specific functions (News, Community, and Memory lane), as well as additional data layers via the What's Happening section (About my area, Planning applications, Funeral notices, Public notices, Corona virus), as well as Additional Services (Local services, Homes near you, Things to do, Items for sale). The website footer indicates that at one point the application included a map of news, though that is no longer active. In fact, there is not currently any sort of spatial representation of news

or data, presuming that any user must already understand the spatial layout of the covered countries and textually search for spatial indicators.

D.2 Manual geo-annotation examples

GeoAnnotator Workbench, developed by (Cai & Tian, 2016), is a response to the variability, specificity, and granularity in local references to places, and an attempt to reconcile the challenge of coding place references in local news documents. The project supports suggestive geocoding, in which toponyms are identified and potential gazetteer matches are presented to a human annotator to disambiguate. The results are ranked by their likelihood of match, using heuristic rules for local geo-annotation (such as previous annotations being more highly favored) to prioritize recommendations. In the event that no appropriate gazetteer entry exists, a human annotator may create a new entry to associate to the place, resulting in a progressively enriched gazetteer. The system uses Google Maps as a global gazetteer and Nominatim, a VGI gazetteer associated with OSM, serving as both a local (with hyper granularity for the county of the local news source) and global gazetteer to form the base from which to grow the hyperlocal gazetteers. The study concluded that 16% of references encountered no ambiguity, 59% required heuristic disambiguity, and 12% required custom footprints (that is to say that these absolutely required human intervention).

GeoAnnotator, a similarly named tool created by (Karimzadeh & MacEachren, 2019), assists with post-publication geo-annotation of event documents. Similarly recognizing that this process requires previous knowledge of the locale or ability to find and associate these references, they have developed a system that identifies and matches toponyms, requiring human confirmation or further disambiguation. Similarly to (Cai & Tian, 2016), the authors recognize the importance of contextualization of the source and story as the results of any such tool are domain-dependent, requiring tuning to the geo-coverage area and level to perform well.

Hozint advertises using artificial intelligence (AI) to support the automatic geotagging of locations mentioned articles so that they can be mapped to point locations, specifically in the application of risk intelligence analysis. Users may input the URL of a published news article, from which a crawler will extract the content text and identify potential locations, recommending more likely focus locations. These can be confirmed, at which point the articles will be available for further analysis via online GIS tools.

D.3 Journalist geo-annotation examples

Several tools exist for associating geolocation to webpages, posts, articles, etc.

GEOmyWP is a paid wordpress plugin developed by (Fitoussi, n.d.) that supports easy geotagging of blog posts and integration of user location to improve a browsing

experience. Users can search for posts proximal to their location or defined areas of interest, or review maps with various post types. The geotags resolve to a point location, defined either by input coordinates, selection on a map, or autocomplete search functionality. The system is based on Google Maps or Leaflet.

Bloom is a project accessible via WordPress plugin or API that allows agencies to geotag their articles (with one primary and up to 30 secondary locations), and provides additional functionality via plugins to integrate spatial features into agency websites, such as nearby searching and contextualization maps. The geotagging application uses point definition of addresses, businesses, or POIs as potential input. Readers can save interests (topics and locations), and sort by location proximity or recent publish dates. There is no feature to associate temporal elements to the content, nor can one explore areas. Bloom is currently being used by several news agencies in the United States, including New York and California.

E

Preliminary Specification

E.1 Público corpora

Elizabeth Fernandes, Head of Audience Insights and Analytics at Público, provided the results of the following query of Público's corpora in January 2021. The general corpora includes 215 articles related to "local" or "lisboa" of 3,701 published in October. This information was used in the informative corpora.

```
SELECT NEWS_TITLE, NEWS_URL
FROM (
    SELECT *
    FROM (
        SELECT *
        FROM `publico-paywall.publico_dw.NEWS_DB`
        WHERE
            NEWS_DATE>'2020-09-30 00:23:00' AND
            NEWS_DATE<'2020-11-1 00:00:00'
    ) as A
    JOIN (
        SELECT *
        FROM `publico-paywall.publico_dw.NEWS_TAGS`
    ) as B ON A.NEWS_ID=B.TAGS_NEWS_ID
)
WHERE
    NEWS_CHANNEL_NAME like ('Local') OR
    NEWS_TITLE like '%Lisboa%'
GROUP BY NEWS_TITLE, NEWS_URL
```

E.2 Geonews portal user story

Insights provided via video interview with Diogo Queiroz de Andrade (a prolific journalist for several major Portuguese newspapers and Creative Director of Observador) and Catarina Carvalho (Founder, Editor, and journalist of A Mensagem) (Queiroz de Andrade & Carvalho, 2020).

Small- to medium-sized newspapers use WordPress with (or other such tools) as the digital platform on which they publish articles. An additional plugin is required to incorporate the main brands of map (Google Maps, Infogram, MapQuest, StoryMaps, etc.).

A Mensagem has interest in leveraging existing geographical data into their news stories, such as the information available in the Lisbon GeoPortal and Maps.Me (OSM).

They would also like to map the events covered by their website with basemap contextualization. However, they are also concerned about exposing area that have not yet been covered (news deserts).

The ability to connect with people in different areas is also of interest. Newsletters and push alerts are the industry standard for this kind of engagement, however users often get frustrated and eventually disable these. Even so, the ability to notify readers of incidents in their proximity or defined areas of interest is intriguing.

The agency already has several journalists in the field and would like to leverage data from the municipality.

An existent digital tool of note is In Your Area, which at the time was communicating COVID cases for sub-city areas in the United Kingdom, and provides relevant information (news stories, demographic data, market places, etc.) to users in the zip code where they resides, as well as additional zip codes of interest.

F

Example article in Apregoar

Consider the following story (translated from its original Portuguese) from a local newspaper in Lisbon, Portugal. Ideally, a geoparser would be able to identify and disambiguate the following event-occurring toponyms: Campo de Ourique, Parada Garden, Baixa Pombalina, and Teófilo Braga Garden. However, the location of the event (the planned station for the metro extension) is Teófilo Braga Garden (with endonym Parada Garden). The event-relevant toponyms are Teófilo Braga Garden, as the planned location of the metro station in question, and Campo de Ourique, the smallest administrative boundary associated with the area. The event-irrelevant toponym is Baixa Pombalina, as it is not relevant to the definition of this particular event (Lee et al., 2019).

> CAMPO DE OURIQUE: Residents want to take advantage of the arrival of the subway to double the size of Parada Garden
>
> A group of about two dozen residents presented a proposal to minimize the impact of the works on the garden and enhance opportunities. They want the garden to be an urban super-island, a space where children "can run around freely". When one looks at Campo de Ourique from satellite images, a patch of green in the center of the parish stands out. There, in the grid pattern of the neighborhood that was designed in the same way as the Baixa Pombalina, there is a block that is a green spot. For those who live in the neighborhood, there is no need to look at maps. The Teófilo Braga Garden, better known as the Parada Garden, is the very center of life in the parish and a kind of oasis in the midst of the dense urban and construction that surrounds it.

The full article also references a new plan for the metro station: in which the garden will be enlarged, overtaking the roads within a block radius on all sides. This new place wouldn't be listed in any pre-existing gazeteer, however the area is well defined textually and pictorally (Figure F) by the author. A well-informed reader can understand (using the text and the accompanying pictures) where the boundaries of the

proposed garden lie. An automatic system may successfully extract all of the toponyms listed as boundaries or POIs included within the proposed area, but may not create an accurate, continuous area in which the referenced roads are bisected in the correct locations. Published on 29 September, 2022, the article references time only sparingly. There is a mention of expected construction of the metro station until the year 2026, and ongoing construction in the garden for a period of two years during that time, but without specific dates associated to the expected disruptions.

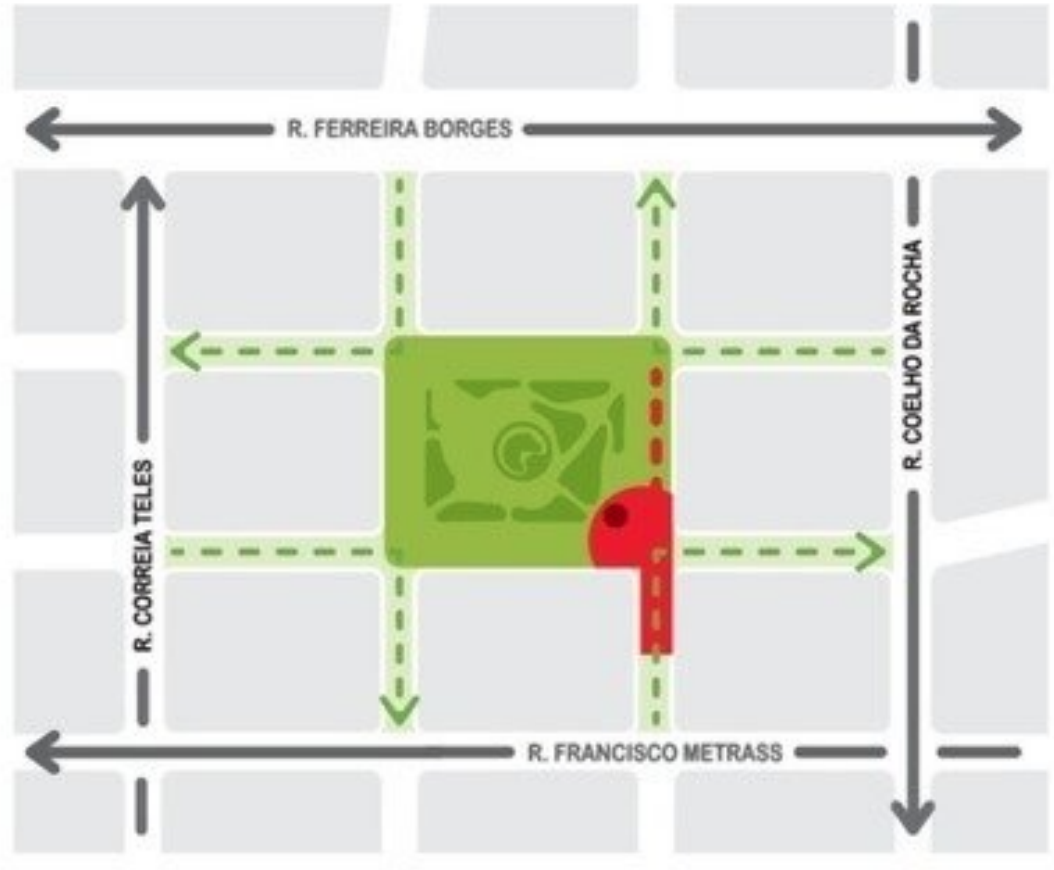

Figure F.1: Proposed expansion of Teófilo Braga Garden in Lisbon, Portugal

F.1 Publisher tool

The process of publishing an article will associate several attributes to it, usually these include story-level values (commonly, those shows in Table F.1 in the "Story" section). In the Apregoar publishing tool, instance values may also be attributes, which also capture the where, when, and descriptions of event instances described within the article. The following table shows these values for the example article elaborated above.

Once published, the journalist may review the article attributes via the interface shown in Figure F.1, including an interactive map showing the distribution of the

instances on a map with additional details.

Attribute	Type	Value
	Story	
Title	Thematic: Text	Campo de Ourique: moradores querem aproveitar chegada do metro para aumentar Jardim da Parada para o dobro
Summary	Thematic: Text	Grupo de cerca de duas dezenas de moradores apresentou uma proposta para minimizar o impacto das obras sobre o jardim e potenciar as oportunidades. Querem que o jardim seja uma superilha urbana, um espaço onde as crianças "possam correr à vontade".
Publish date	Temporal : Date	2022-09-29
Web link	Thematic: Text	https://amensagem.pt/2022/09/29/campo-de-ourique-obras-metro-lisboa-arvores-jardim-da-parada-superilhas-crescer-dobro-aumentar/
Section	Thematic: Text	Bairros
Tags	Thematic: Text	comunidade, mobilidade,Espaço público
Author	Thematic: Text	Frederico Raposo
Publication	Thematic: Text	A Mensagem
	Instance 1	
EGazetteer: campo de ourique		
Place name	Thematic: Text	Campo de Ourique
Description	Thematic: Text	Freguesia onde o metro vai ser inserido
Temporal type	Thematic: Text	Persistent
Begin time	Temporal: Datetime	[null]
End time	Temporal: Datetime	[null]
Description	Thematic: Text	Continual
	Instance 2	
OSM Gazetteer: Baixa, Lisbon		
Place name	Thematic: Text	Baixa Pombalina
Description	Thematic: Text	O quadriculado do bairro semelhante ao Campo do Ourique
Temporal type	Thematic: Text	Persistent

`Begin time`	Temporal: Datetime	[null]
`End time`	Temporal: Datetime	[null]
`Description`	Thematic: Text	Desenhado antes do Campo de Ourique
	Instance 3	
Custom footprint		
`Place name`	Thematic: Text	Superilha urbana do Jardim da Parada, projetada
`Description`	Thematic: Text	A expansão do jardim com o metro novo, proponha por a comunidade, com acceso pedonal
`Temporal type`	Thematic: Text	Persistent
`Begin time`	Temporal: Datetime	[null]
`End time`	Temporal: Datetime	[null]
`Description`	Thematic: Text	Prevista para implementação com o novo metro
	Instance 4	
EGazetteer: jardim téofilo de braga		
`Place name`	Thematic: Text	Jardim da Parada
`Description`	Thematic: Text	Lugar existente do jardim
`Temporal type`	Thematic: Text	Persistent
`Begin time`	Temporal: Datetime	[null]
`End time`	Temporal: Datetime	[null]
`Description`	Thematic: Text	Já existente

Table F.1: Manually extracted attributes

F.2 Explore tool

A reader or researcher may use the explore tool to search the Apregoar database (all sources) to identify articles of interest. To filter the results, the user may open a panel in which they may define story level filters (dates of publish, tags, sections, sources, and/or authors) and instance level filters (instance dates, type of temporal definition, type of spatial definition, areas that fall within administrative boundaries). The user may also use a search feature to identify places, or draw areas to which the filtered events should correspond (with four types of specific spatial relationships as options). Figure F.2 shows the filter panel with a publish date range and polygon spatial search fields activated.

The web app performs a query of the spatial database and returns the search results in the Explore GUI. The top row of results represents the returned stories, and just

Apregoar Dashboard Publicação Jornais Geradores Exploradores Profile

Campo de Ourique: moradores querem aproveitar chegada do metro para aumentar Jardim da Parada para o dobro

Bairros, 2022-09-29
A Mensagem
Grupo de cerca de duas dezenas de moradores apresentou uma proposta para minimizar o impacto das obras sobre o jardim e potenciar as oportunidades. Querem que o jardim seja uma superilha urbana, um espaço onde as crianças "possam correr à vontade".

Espaço público, mobilidade, comunidade

Instâncias

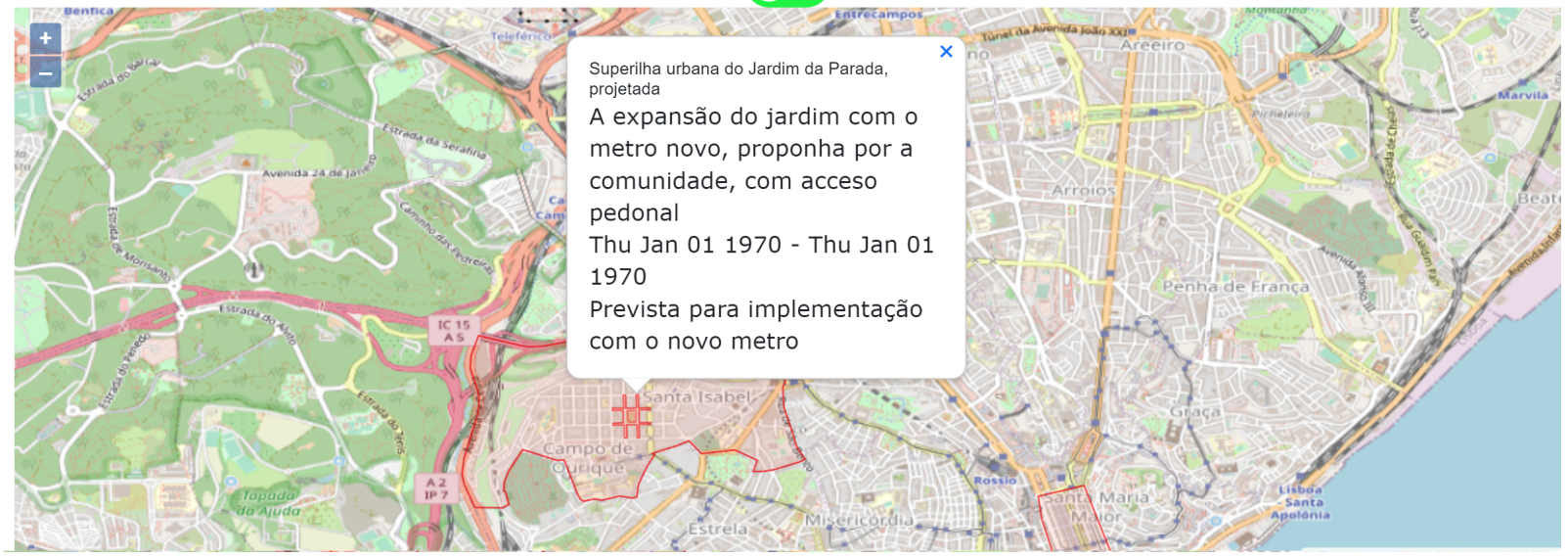

Figure F.2: GUI of a "published" article from the agency's backoffice

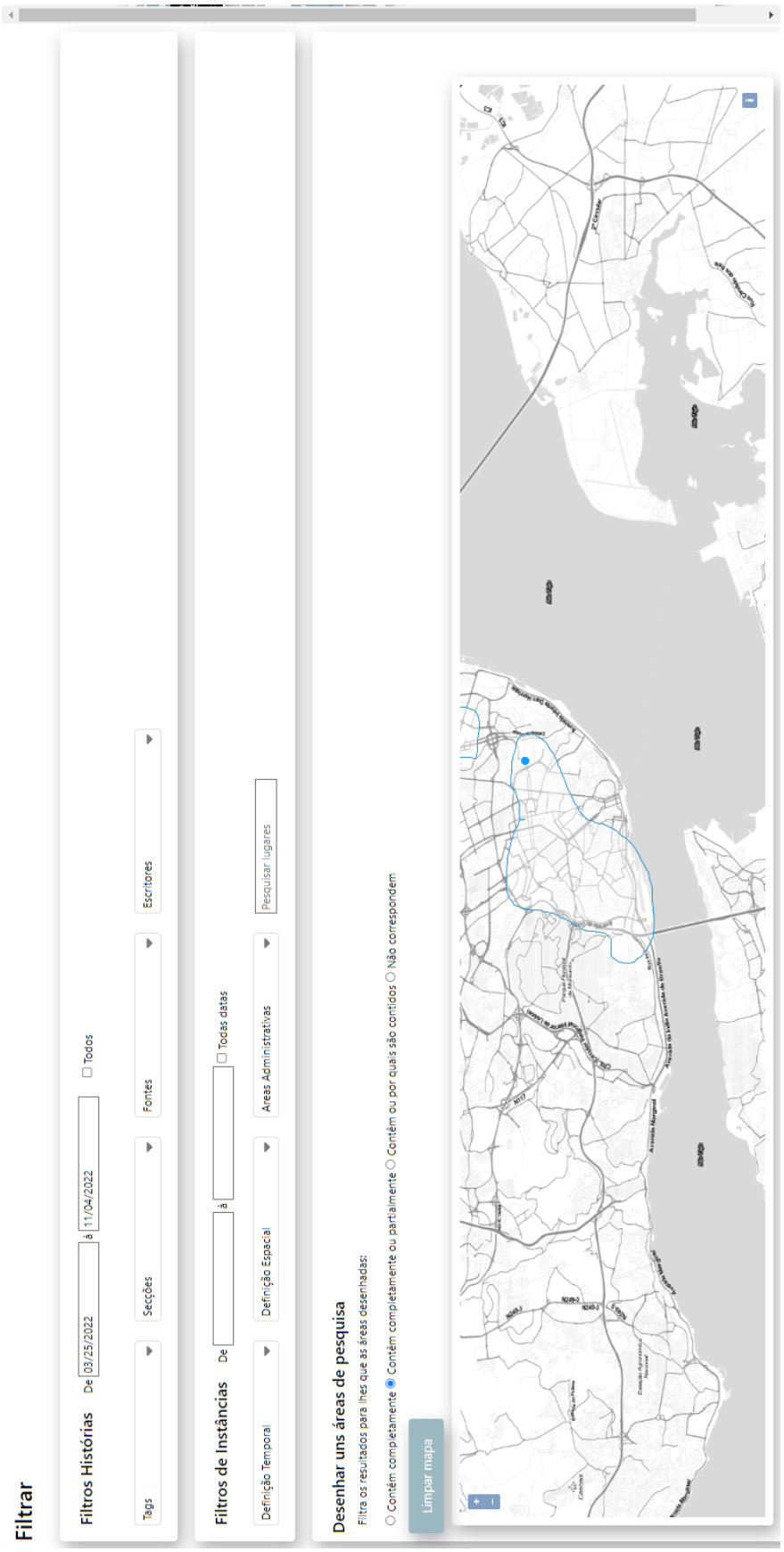

Figure F.3: GUI of filtering functionality in the Explore tool

below all of the returned instances are displayed. The map below these lists shows all of the footprints of the associated instances. See Figure F.2.

When a story or instance is selected, it list and map elements as well as all other elements pertaining to the same story, are highlighted to various degrees: the element selected is brightlighted (outlined in a bright yellow to draw the most focus), while the related elements are highlighted (outlined in blue). All unrelated items (returned search elements that do not pertain to the identified story) are nolighted (greyed out to detract focus). Any instances associated with the identified story that were not returned as search results are lowlighted (outlined in a light yellow to indicate presence without drawing focus). See Figure F.2.

When an incident or article is selected, the user may scroll down to see the details of that article. Story attributes are situated at the top, and instances appear as block elements with their own attributes below. These may be clicked to highlight the specific instance of interest on the map. This details area also provides a link to the original article (hosted by the original source), as well as a simulated "original", which includes the article with the context map to demonstrate that functionality. See Figure F.2.

F.3 Context tool

Figure F.3 shows a simulated version of an "original" article with a map representation via the Context GUI to support contextualization, as well as options to spatially explore other stories belonging to the same source, as the tool may be integrated into agency articles, potentially via the anticipated WordPress plugin.

Figure F.4: GUI of filtered results in the Explore tool

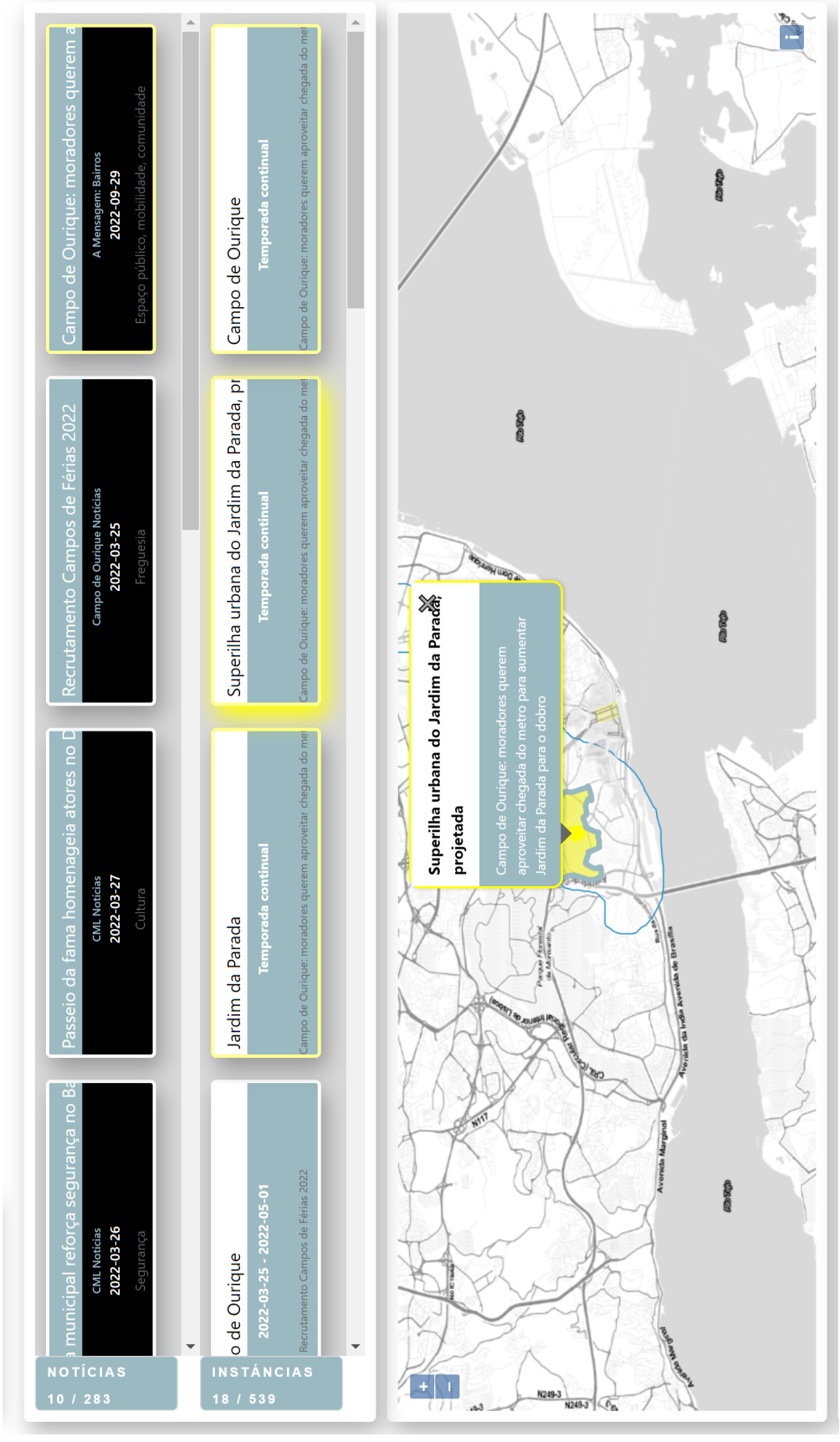

Figure F.5: Highlighting of story and instance selection in the Explore tool

Figure F.6: Detailed view of story and instance selection in the Explore tool

Figure F.7: GUI of simulated original article with contextualization map

G

Apregoar queries

G.1 Geonotícia query

```
CREATE OR REPLACE VIEW apregoar.geonoticias AS
SELECT stories.s_id, instances.all_gaz, stories.title, stories.summary,
stories.pub_date, stories.web_link, stories.section_name AS section,
stories.section_id, stories.tag_name AS tags, stories.tag_id,
stories.author_name AS author, stories.author_id, stories.publication_name
AS publication, stories.publication_id, stories.u_id, instances.i_id,
instances.t_begin, instances.t_end, instances.t_type, instances.t_desc,
instances.p_name, instances.p_desc, instances.num_egaz, instances.e_ids,
instances.e_names, instances.p_id, st_area(instances.all_gaz) AS area,
st_astext(st_centroid(instances.all_gaz)) AS centroid_text
FROM (
    SELECT stories_1.s_id, stories_1.title, stories_1.summary,
    stories_1.pub_date, stories_1.web_link, stories_1.u_id,
    array_agg(DISTINCT publications.publication_id) AS publication_id,
    string_agg(DISTINCT publications.publication_name,',') AS
    publication_name, array_agg(DISTINCT tags.tag_id) AS tag_id,
    string_agg(DISTINCT tags.tag_name,',') AS tag_name,
    array_agg(DISTINCT sections.section_id) AS section_id,
    string_agg(DISTINCT sections.section_name,',') AS section_name,
    array_agg(DISTINCT authors.author_id) AS author_id,
    string_agg(DISTINCT authors.author_name,',') AS author_name
    FROM apregoar.stories stories_1
    LEFT JOIN apregoar.publicationing ON stories_1.s_id =
    publicationing.story_id
    LEFT JOIN apregoar.publications ON publicationing.p_id =
    publications.publication_id
    LEFT JOIN apregoar.tagging ON stories_1.s_id = tagging.story_id
```

```
    LEFT JOIN apregoar.tags ON tagging.t_id = tags.tag_id
    LEFT JOIN (
SELECT sectioning.s_id AS st_id, sectioning.story_id
        FROM apregoar.sectioning
) sects ON stories_1.s_id = sects.story_id
    LEFT JOIN apregoar.sections ON sects.st_id = sections.section_id
    LEFT JOIN apregoar.authoring ON stories_1.s_id = authoring.story_id
    LEFT JOIN apregoar.authors ON authoring.a_id = authors.author_id
GROUP BY s_id
) stories(s_id, title, summary, pub_date, web_link, u_id, publication_id,
publication_name, tag_id, tag_name, section_id, section_name, author_id,
author_name)
LEFT JOIN (
    SELECT inst_e.i_id, inst_e.t_begin, inst_e.t_end, inst_e.t_type,
    inst_e.t_desc,         inst_e.p_name, inst_e.p_desc, inst_e.s_id,
    inst_e.u_id, inst_e.num_egaz, inst_e.e_ids, inst_e.e_names,
    inst_e.egeom, ugaz.p_id, ugaz.geom AS ugeom,
    st_collect(ARRAY[ugaz.geom, inst_e.egeom, ingaz.ngeom]) AS all_gaz
    FROM (
        SELECT inst.i_id, inst.t_begin, inst.t_end, inst.t_type, inst.t_desc,
        inst.p_name, inst.p_desc, inst.s_id, inst.u_id, egaz.num_egaz,
        egaz.e_ids, egaz.enames AS e_names, egaz.egeom
        FROM apregoar.instances inst
        LEFT JOIN (
            SELECT inst_egaz.i_id AS instance_id, count(inst_egaz.i_id) AS
            num_egaz, array_agg(inst_egaz.e_id) AS e_ids,
            string_agg(agaz.name, ','::text) AS enames,
            st_collect(array_agg(agaz.geom)) AS egeom
            FROM apregoar.instance_egaz inst_egaz
            LEFT JOIN apregoar.egazetteer agaz ON inst_egaz.e_id = agaz.e_id
            GROUP BY inst_egaz.i_id
        ) egaz ON inst.i_id = egaz.instance_id
    ) inst_e
    LEFT JOIN (
SELECT iugaz.i_id, array_agg(DISTINCT ug.p_id) AS p_id,
st_collect(ug.geom) AS geom
        FROM apregoar.instance_ugaz iugaz
        LEFT JOIN apregoar.ugazetteer ug ON iugaz.p_id = ug.p_id
GROUP BY iugaz.i_id
) ugaz ON inst_e.i_id = ugaz.i_id
    LEFT JOIN (
```

```
        SELECT inst_ngaz.i_id, st_collect(array_agg(ngazetteer.geom)) AS ngeom
        FROM apregoar.instance_ngaz inst_ngaz
        LEFT JOIN apregoar.ngazetteer ON inst_ngaz.n_id = ngazetteer.n_id
        GROUP BY inst_ngaz.i_id
    ) ingaz ON inst_e.i_id = ingaz.i_id
    ORDER BY inst_e.i_id
) instances ON stories.s_id = instances.s_id
ORDER BY stories.s_id;
```

G.2 Ugazetteer access query

```
CREATE OR REPLACE VIEW apregoar.access_ugaz AS
SELECT uugaz.p_id, uugaz.p_name, uugaz.p_desc, uugaz.geom,
    uugaz.ug_created, uugaz.ug_edited, uugaz.u_id, uugaz.publication,
    inst2.s_id
FROM (
    SELECT ugaz.p_id, ugaz.p_name, ugaz.p_desc, ugaz.geom,
        ugaz.created AS ug_created, ugaz.edited AS ug_edited,
        ugaz.u_id, users.organization AS publication
    FROM apregoar.ugazetteer ugaz
    LEFT JOIN apregoar.users users ON ugaz.u_id = users.u_id
) uugaz
LEFT JOIN (
    SELECT inst.s_id, iugaz.p_id
    FROM apregoar.instance_ugaz iugaz
    LEFT JOIN apregoar.instances inst ON iugaz.i_id = inst.i_id
) inst2 ON uugaz.p_id = inst2.p_id;
```

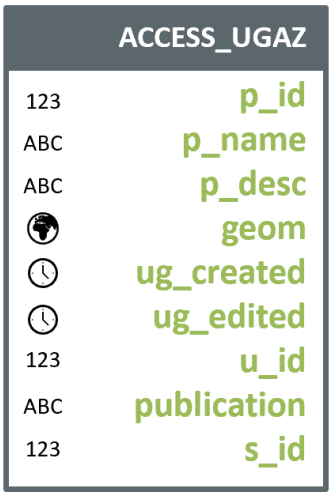

Figure G.1: Ugazetteer access query

G.3 Egazetteer filter query

```
CREATE OR REPLACE VIEW apregoar.egaz_filter AS
WITH u_egaz AS (
    SELECT sa.e_id, count(sa.p_id) AS count_u
    FROM (
        SELECT DISTINCT spatial_assoc.e_id, spatial_assoc.p_id
        FROM apregoar.spatial_assoc
        GROUP BY spatial_assoc.e_id, spatial_assoc.p_id
    ) sa
    GROUP BY sa.e_id
), e_egaz AS (
    SELECT instance_egaz.e_id, count(instance_egaz.i_id) AS count_e
    FROM apregoar.instance_egaz
    GROUP BY instance_egaz.e_id
)
SELECT e.e_id, e.total_count, lower(egaz.name) AS e_name
FROM (
    SELECT COALESCE(u_egaz.e_id, e_egaz.e_id) AS e_id,
        COALESCE(u_egaz.count_u, 0::bigint) + COALESCE(e_egaz.count_e,
        0::bigint) AS total_count
    FROM u_egaz
    FULL JOIN e_egaz ON u_egaz.e_id = e_egaz.e_id
    ORDER BY (COALESCE(u_egaz.e_id, e_egaz.e_id))
) e
LEFT JOIN apregoar.egazetteer egaz ON e.e_id = egaz.e_id;
```

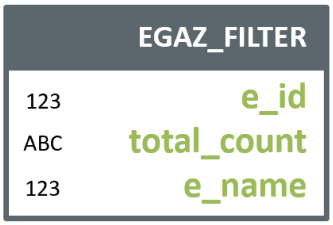

Figure G.2: Egazetteer count query

UNIGIS PT

NOVA IMS Information Management School

2022	APREGOAR: Development of a geospatial database applied to local news in Lisbon	Caroline Gilman Wentling