IN PROGRESS
An open source spatial news web app development project
Masters in GIS&S Thesis

Callie Wentling
Project advisor: Professor Doutor Marco Painho, Ph.D

August 26, 2021

Project summary

This project seeks to develop a proof-of-concept (POC) open source (OS) and free software (FS) web application (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. This geospatial element is expected to provide an additional dimension of understanding that allows users to better contextualize news stories, search repositories, or monitor spatial/temporal trends at a community level (within a city). 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.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Context

The ongoing COVID pandemic has highlighted the value of the visualization of information on a map, not only for specialists to monitor and predict viral outbreaks, but to arm the public with empowering information as well. Of course, the value of geographic information systems (GIS) goes beyond public health services 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 (SA), temporal (TA), and thematic attributes (ThA). For example, a user on AirBnB may filter all apartments with high speed WiFi (ThAs) available in the Estrela neighborhood and within walking distance to a market (SAs) from Aug 1 to Aug 7, 2020 (TA).

Yet, though this type of manipulation is commonplace in the 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. 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. 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 these cases, as well as a host of others, there is obvious disconnect between the existence of data and its usability. Though many search engine queries contain geographic keywords [1], news media enterprises have not yet accommodated such spatial associations to their articles that would provide an expected improved user experience and therefore competative 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 Objectives

This project seeks to develop a set of functional tools that supports the creation and management of a spatial database of local news stories, a publishing interface (associating place and adding records to the database), a user interface (list and map format search, filter and visualization of results from the database), as well as a story visualization plugin (a map displaying the distribution of a story in a contextual map per story page), see Appendix C. This proof of concept (POC) functionality should demonstrate the value of new spatial products in news media, and provide a basis from which meaningful projects may be developed for mass media applications in the future.

1.3 Structure

2 Literature review

2.1 Spatial information

"Location is involved with everything" [2]. Location is the definition of a point in space relative to the earth's surface. It describes, along with thematic and possibly temporal attributes, phenomena, such as where events happen or where something is spatially situated [3]. Location can be defined in a variety of ways, though perhas most often interpretted by humans in two dimensional coordinates associated with a particular spatial reference system (SRS). A common example is the often used longitude and latitude coordiantes 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 [4]. Such formal definitions of location allow the connection of disparate phenomena or datasets by providing a common framework within which they can be compared [2].

2.1.1 Geographic information systems

Though often relevant for decision making at the personal, organization, and regional levels, spatial data is underutilized in formal analysis [2].

Geographic information systems (GIS) are computer programs that suppor the collection, sharing, processing and visualiation of geospatial data and its resulting information [5]. This georeferencing (associating data to a map) is fundamental for understanding where people, things, and events are, were, or may be [6,7]. The resulting geogrpahical datasets, in the forms of maps and features, provide an opportunity to orient collaborators, share experiences, and challenge presumptions of the users [5,8].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 [2, 7]. It is no surprise, then, that location data is already considered valuable and increasingly being incorporated into at all scales of community operations [2, 9]. One of its most valuable characteristics is the ability to display relational values across time and space [5]. The convolution of GIS technology, the availability of GPS infrastructure, and machine learning techniques promote a myriad of real-time and spatial services with research, commercial, and security implications [10, 11]

2.1.2 Feature definition

Where something occurs is stored in raster (pixelated image) or vector (feature) format. 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 (incoporating 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 [12]. 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 [13]. The choice of feature type affect both the analytical processing as well as the deducations made when viewing the results, the latter being further affectedb y placement and marker represen-

tation [12]. Ultimately, the application and requirements of the 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 vector type.

A place, by contrast to a location, is a geospatial definition, and potentially much more. They include the objective, such as descriptions of the objects physically present, such as a river or a building, but there are other elements to consider. Places are subject to the multiple identies strewn upon them by those who experience them under different conditions, formed by direct engagement or passed anecdotes [14]. 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 of the definer. Place, then, can have history or nostalgia, and these informal relationships may be constantly morphing [9]. This is especially apparent when considering a microcosm such as a neighborhood, community, area as a place. The experience of those who live, work, and visit a place and constantly being fedback into these areas, leveraging local knowledge and expectations of what the place is [15]. For all of these reasons, it can be challenging to associate thematic attributes 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 [13].

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 [13, 14]. More likely, the data on either side of the line are heterogeneous, more fluidly changing state over a shore of values, rather than a counterposition the other on the opposite side of the enclosure 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 [6].Therefore, any static representation is already a distortion of reality (just as or more than any data representation is an abstraction) [5]

One tool to address these fluid definitions is neogeography, a new way of understanding place by combining location based services (LBS) and volunteer geographic information (VGI) [16]. 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. Many of the same tools that allow us to remotely connect with each other include geolocation, making these social media users producers of vast quanitites of spatial data. This can be used to identify new, digitally relevant points of interest of areas where people convene, or be leveraged as expert data when users describe their own communities [9]. They are also, quite convenently, composed of the building blocks of geographic data [3].

2.1.3 GeoIntelligence

2.1.4 Geoportals/mashups

2.1.5 Commercial geospatial platforms

2.1.6 Location extraction from text

2.1.7 Event extraction

2.1.8 Gazetteers

2.1.9 Local lexicons

2.2 Citizen empowerment

2.2.1 Smart communities

Though most often referred to on the city level, smart communities leverage information and communication technologies (ICT) (networks) and various sources of data (sensors) to address and improve the functional needs of its population (actuators), engaging its "users" to develop citizen-centered interventions and responding to their changing needs [9,17,18]. In fact, "[a]n active and engaged citizen is indeed the main driving force of a 'smart city"' [19]. 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 by affecting goverments planning procedures and services [17]. Beyond efficiency, citizens require safer, more enjoyable living experiences in all aspects of their lives. Governments may accommodate the public interest [18] by incorporating four dimensions (intelligence, digital, open, and live, referring to its social and informational infrastrucutres, open governance, and continuity of adaptation, respectively) of smart communities [19]. The identification and monitoring of community dynamics requires "sensing life" though open dialogues with constituents as well as internet of things (IOT: the integration of networked hardware sensors to monitor and/or interact with their surroundings) technology [9]. This sensing infrastructure leverages various sources of data to determine the state of various subsystems and support interventions for improvement. Ideally, it may 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 [8].

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" [9] has an internal structure [14], with corresponding spheres of influence of its nodes within and outside of these. As quickly as informational technology (IT) tools provide new means of characterising the immediate, physical geographic area of a community node, it also supports 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" [14], 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 reponses to specific events" [19].Data in general is already highly regarded as a key comodity for developing an economy [20], To harness the value of this ever-expanding resource, community operations should accommodate methods for capture, exploring, and sharing this data, spatial or otherwise, and its processed results [9].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 participants of the community [7]. 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 [18]. Members of such knoweldge societies [21] , investigators and entrepreneurs or anyone with with access to technology, are better equiped to address local psychographics ("the prevailing interests of people in an area" [22] in nontraditional or niche applications [5]. 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.

2.2.2 Communication

In the course of its operations, a smart community should facilitate a "shared understanding of what is happening" within it [21],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. Phsyical 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 networkng platforms [23]. Following suit, many news channels and government communication departments have incorporated digital distribution strategies, often that leverage social media to engage readers and direct traffic to their channel paltforms. This allows not just local eyes on local announcements, but also invites remote viewers to participate [23].

Stemming from the assumption that "storytelling is the most effective way to merge meaning and emotions" [24], a tremendous and increasingly more ubiquitous tool for effective and relatable communication is data visualization [20, 25]. Data visualization products and inclusions have migrated beyond the niche tech or empresarial applications to "a part of the fabric that is modern culture", threading their way into newspapers, fasion lines and books [26]. Studies indicate that readers prefer pictoral and summarial forms of information (as opposed to purely textual) [23]. Visuals can provide additional context, identify changes, reveal patterns, and showing and distinguishing between relationship types [27], ultimately "connect[ing] numbers to what they really stand for: knowledge, behaviors, people" [20]. Users, whether they be the general public or decision makers, are expected to have some data visualization literacy (DVL) [28]. This mutual expectation of information producers, consumers, and actors to present and ingest efective representationsf is reenforcing its importance and creating new standards of competencies. "Every publisher and journalists knows the value of charts and wants more of them" [29], 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 [25, 26]. Simultaneously, academia has established a variety of digital visual literacy frameworks (DVL-FW) 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 [28]. 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 increasingly being employed to answer *where* and related questions as they are more easily interpreted and

remembered [28]. They illucidate spatial relationships using layers of data contextualized by basemaps (such as raster images or vector representations) [5, 8]. Especially when presented digitally, online 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. Especially when evaluating foreign areas, maps can provide especially valuable context by concisely representing proximities and directional situations, versus relying on verbal descriptions that may perhaps be more easily misconstrued [27].

In any case, data should be used and interpreted cautiously. Data records are an abstraction of the real world [20]. Often, visualization of data disclude elements of uncertainty [26] or are developed prematurely (without proper analysis) or improperly (misleadingly) [28,30]. Further, visualization designer 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 [20, 28, 31].

2.2.3 Public participation

Public participation is a critical element of citizen empowerment. democratic vibrancy, and innovation [18].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 [13]. Further, participation strengthens a community by building social capital amongst its participants, demonstrating trust between members [23]. High forms of civic engagement (CE) assume that citizens have the power to influence decisions that will touch their own lives, whether through active dialogues or other means of engagement. Thoug 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 organiation and the community as a whole. Updated strategies, especially those that include presentially and digitally hybrid participation options, may engage larger audiences, facilitating greater particpiation while mitigate possible digital divides in participating demographics [18,23]. It can also prolong interactions, allowing all stakeholders to reevaluate options and motivations throughout the entire process [18]. These services may be government or institutional services (ex: Lisboa Participativa), non-institutional platforms (CitySourced), or commercial products leveraged for engagement (ex: NextDoor).

As an "inherently spatial" element, public participation should not be disconnected from this dimension [13]. Spatial information is critical for making educated decisions on key human issues [7].Though this clearly applies to decisionmakers in their respective fields, access to location services is must be a given for all modern society [7]. "[A] city could not be smart without spatially enabled citizens" [9], 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 Infromaiton and Analysis in the United States of America which establisehd the public particiaption geogrpahic information system (PPGIS) to better accommodate marginalized populations [12]. It can be especially powerful to visualize the impact of interventions of underrepresented communities at scale [**?**]. At its core, a PPGIS represents an abstract of thematically interesting features, contributing to a communal understanding of place [12]. "[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 [16]. 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" [32].From this, such online tools should include elements of understanding the decision making processes and tracking its progress, both of which support transparancy, as well as opportunities to influence it, such connection, sharing of information, a platform for developing ideas [18].

A critical element of any spatial understanding, smart or participative, is the collection of data with a geospatial element. Beyond intermittent and reprentative 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 [9]. Whether active or passive – describing whether the data generation is consciously initiated by a participant (such as participating in a forum) or collected in the backgroud of regular activity (such as via a smart phone app), both of which should be intentionally shared if accessed by a third party – 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), volunteered geographic information (VGI) is critical to the understanding of "citizens' social synergies in the urban context" [13, 23]. 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 [13], 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 [16].

- "But more important for a smart city is its capability to capture the sense of places. A city is not a machine, bur rather made by people local actions and feelings." [19].

-Distinguishing characteristics of local knowledge: "It is based on experience. It is developed over time by people living in a given community, and is continuously developing. It is embedded in community practices, institutions, relationships, and rituals. It is held by individuals or communities. It is dynamic and changing. Based on these characteristics, we may anticipate that local knowledge is unique from place to place. Therefore, the gazetteer used for geo-referencing local newspaper articles should be place-specific." [15]

2.3 News localization

2.3.1 News evolution

2.3.2 Personalization

2.3.3 Blogs

2.3.4 Indie journalism

2.3.5 Local journalism

2.3.6 Georeferenced News

2.3.7 Automated location extraction from published corpora

2.3.8 Manual location extraction from published corpora

3 Apregoar: A Spatial News WebApp

3.1 Concept

Apregoar is a foundational, proof of concept web application that seeks to demonstrate the possibilities of intentionally associated temporal and spatial attributes to news stories by media publishers for an improved user experience for traditional readership, as well as improved searching capabilities for researchers and informational dashboards for monitors. Through this portal, a variety of users may interact with news media in new ways to glean additional insights about relevant places, histories, or previously obscurred geospatial patterns. Each of these 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 will, therefore, focus on only a subset of these to demonstrate the concept. The conclusion of this demonstrative version will serve as a draft for initial testing by a variety of users, draw constructive criticism, and ultimately continue to evolve into the full featured platform that it has the potential to become. As such, the full concept is desribed here, while methodology will outline specifically work done under the scope of this project to develop the Apregoar WebApp.

It is expected that the association of specific place (potentially non-conforming to existing administrative boundaries or defined points of interest) 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 requiring adjustment of publishers' processes to maintain, may 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, it could inspire a shift of the industry standard towards integration of spatial attributes and spatially related products. 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 [8].

3.2 Views

3.2.1 Georeferenced Data Entry

Publishers interested in harnessing the foreseen spatial benefits of geospatial news for their own research and readership experience may incorporate the georeferencing step into their publishing process. Though ideally developed in the future as a plugin to existing publishing processes (such as a wordpress plugin for organizations leveraging the NewsRoom tool), initially the corpora will be added separately through the portal itself. At this stage, publishers will enter standard article information such as title, summary, web-link to their hosted article, define a story type (report, article, editorial, etc.), section, and tags. Additionally, they will be prompted to associate one or more instances: temporal/spatial definitions of the events described within the article. The temporal element will be defined to the day, requiring a start and end day (which may be the same), as well as the

opportunity to textually describe the interval as well. To georeference the incident, they will have the opportunity to draw one or multiple polygons describing the area where the events occured. If appropriate, the publisher may define multiple incidents, if multiple events occured or they are describing impacts that changed over time.

In future variation, publishers should be able to "whitelabel" their entries, creating a platform that they can integrate into their own websites that include only their own publications, 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.

3.2.2 Spatial News Database

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. 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, or any personal preferences or history users may elect to set.

3.2.3 Contextualization Maps

It is understood that the integration of visualization into existing products is an important part of relevance and commercial value propositions [26]. Therefore, a key view 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 publications readership. In this view, a page dedicated to a single news event is loaded, with critical data such as title, summary, publication date, author, etc. along with a map situating the associated georeferrenced incidents. Users may use the basemap points of interest or scrolling functionality to oriente themselves to the location of the stories events.

In future iterations, nearby events may also appear in a different color to indicate recommended and spatially relevant stories.

3.2.4 Searching Maps

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 locales. Other readers, however may be more selective and particularly interested in things that are occuring near 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 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 particular name of the occurance. As of yet, news platforms don't integrate spatial searching beyond the incorpoation 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 colloqual titles that the researcher is not privvy 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 thematic attributes) 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 section of existing applications.

3.2.5 Monitoring Dashboards

Though not yet implemented, monitoring dashboards are one of the most interesting potential applications of such georeferenced data. These provide the opportunity to layer geointelligence 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 ground tuth, especially density estimation and analysis of spatial distribution. Perhaps, this can also eventually leverage other interesting infromation, such as demographic, crime, or weather data as potential additional features. However, other interesing analytical methodologies should leverage the data extraction functionality of the tool (future integration of download or API connections) to layer this georeferenced communication data into systems more appropriate to handle this kind of data manipulation and yeild additional geointelligence insights.

3.3 Nuanced place

Place is hard to pin down. Geographical cue words such as ('city of Lisbon', 'just outside of Lisbon', 'Lisbon-based', and 'Rio Tejo' , adapted from [33]) don't have precise spatial definitions, and can queue different understandings in various contects. "Just outside of Lisbon" is a particular area near Lisbon, but likely not a buffer of the city, more likely its an area to the side. Rio Tejo shouldn't necessarily associate a point to the center of the entire of a river flowing thorugh the entire country, but is rather relative based on the context of the story/place (an area of river just off the coast of a particular place, for example). In these cases, there isn't 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 [34].

Where something is happening may be best associated with existing gazetteers of information. These are expected to be point data (addresses, POIs, etc.) or polygons (administrative boundaries). There is a value to leaving the data as is (premitting the

publisher to define a point or a polygon), that can be transformed on the fly as necessary [12]. However, as nothing happens in a location of zero dimensions, the tool gently pushes the user to define a polygon, that provides a consistent experience and can still be transformed as necessary. 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 requiring custom areas (which can use existing boundaries as templates from which to draw) it also pushes publishers 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 [5].

Just as a reader's lexicon develops over time, so should the associated gazetteer. Manual specification allows the user the opportunity to name/address and define these flowing areas as their and general understanding of an event's location changes over time. A military base may change and move over time, the same named thing within a city could come to mean different locations at different times. The flexibility to adjust on the fly is critical to the success of this type of initiative. [33]. These custom designed 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 occured 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 placemaking within the study area, external to any official news communication.

3.4 Participative communities

The tool can be considered a PPGIS system in that those assigning spatial definitions will not be trained GIS users. Rather, they will mostly be composed of jouranlists and publishers and officials or public institutions, though an even more broad definition of "public" is posible as non-institutionally affiliated users are also welcome to contribute their own georeferenced stories. It can also be considered a PPGIS as it will required the assigned 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) [12]. The visualization of this data is then available to any user for further exploration at will.

Its greatest value, however, is as an input to a PPGIS, helping to form a general spatial understanding of the public by providing additional insights to citizens about the areas that affect them – where they live, work, play, or have an interest. 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 (less structured and potentially long form versions of "Na Minha Rua") to gain further citizen feeling or commentary about their physical surroundings.

3.4.1 Distinction from previous work

Unlike many of the example provided in the literature, this effort seeks to georeferrence articles instead of social media posts, most noteably "tweets". Tweets or other specific

data types include structured organizations and data APIs from which automated programs attemp to derive sentiment and/or relation to particular events [35].

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 (news papers, reports, etc.). The association of place to news sources can be studied for its influence on public oppinion and determien how the spread of news affects public oppinion.

This project also fills a different niche than the projects attempting to parse a variety of information (spatial, temporal, and thematica such as sentiment, volatility, etc.) from international jornals 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 human-in-the-loop association, this project seeks to develop a novel standard for allowing a publisher to explicitly assign temporal and spatial attributes 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, avoid misassociation of time and place (such associating activity from Lisbon, Portugal to Lisbon, Ohio in the United States of America, or accurately differentiating between the Distrito de Lisboa, the Município de Lisboa, and the Área Metropolitano de Lisboa.

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 grande 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 placemaking 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 level), the name of place will be understood differently. Moreover, neighbors within a particular parish may define the same neighorhood in different ways, or use different language to describe it. This process allows publishers with a clear understanding of where an event (as the subject of his or her article) is occuring (or was or will occur) to define its boundaries outside of common understandings or administrative definitions. This provides a common defintion of place that can be visually undertsood, and persist beyond changing borders or evolving names.

4 Methodology

4.1 Study area: Lisbon, Portugal

Lisbon is the capital city of Portugal, located close in the central west area of the continental country, just North of the River Tejo. Half a million people reside within the municipality of Lisbon, a total of 2.9 million in the Área Metropolitano de Lisboa (a Nomenclature of Territorial Units for Statistics NUTS II region [36] of 17 municipalities both north and south of the River Tejo), and 2.3 million in the overlapping Distrito de Lisboa (16 municipalities, entirely North of the river) [37].

IMAGE OF DISTRICT VS AML VS MUNICIPIO

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 availabe exlusively online. A Mensagem, for example, is a newer addition 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 Câmara Municipal de Lisboa solicits regular boletins and notícias, as do many of the juntas de freguesias 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 [38]. Portugal more generally also has taken an interest in innovative journalism; João Palmeiro, president fo the 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 [29].

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.

4.2 Requirements

The Apregoar system architecture reflects the needs of the browsers, researchers, monitors, and publishers 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 views will only require reading of selected database records. This should include text, date, and spatial types.
- The system should be implemented with no or low cost maintenance.
- The system should support open source applictions, and therefore utilize openly licenced tools as much as possible.

- The system targets non-professional GIS users, and therefore should be straightforward and easy to use.
- The system will use spatial, thematic, and temporal filtering, and will therefore leverage dataframes 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, which will therefore saved as a or multiple polygon type records, as well as for defining spatial search areas for use in filtering georeferenced articles.
- The main search results will be in map and list formats. Therefore, the system must support this type of data viewing.

4.3 System architecture

The Apregoar platform components were selected based on the above requirements, as well as the functionality described in the previous section. Also considered were the functionality of potential tool, available official and community resources, as well as prior experience in the options.

An event model requires spatiotemporal and attribute elements, with the former requriing specific formats and processing for data manipulation and storage. For this reason, the data storage leverages PostgreSQL with a Postgis extension for its vector data storage capabilities, support in literature ([2,19,39,40]), support for remote connection via python using related packages, and personal experience with the platform.

Connecting the database to the web application is an Apache server for its ease of implementation. For serving spatial features, Geoserver was selected for its vector data support, its documentation, support in literature ([2, 8, 40]) and prior experience.

The backend is based on a the Flask web development platform, chosen for its lightweight implementation and python programming language. Key libraries include Shapely, Geopandas, SQLalchemy, geoalchemy2, and Gdal.

For mapping visualization in the front, Open Layer and Map Box were both considered. Open Layers better adheres to open source standards, though to streamline the proof-of-concept development phase, MapBox was implemented due to its prior experience, as well as integration of a variety of functionality via its javascript library that supports many of the operational, yet not focal, functionality of this project.

Any data preparation or local testing utilized QGIS, also for its adherence to open standards, prior experience, and literature support ([40]).

4.4 Relational data model

The data relates potentially multiple spatial definitions to a particular news story via the association with instances, which layer back in the temporal element of each story. Users are also defined such that published articles can be edited by those who created those records, and then associated with activity of that user (or set of affiliated users, if applicable). This relationship is outlined in Figure 7.

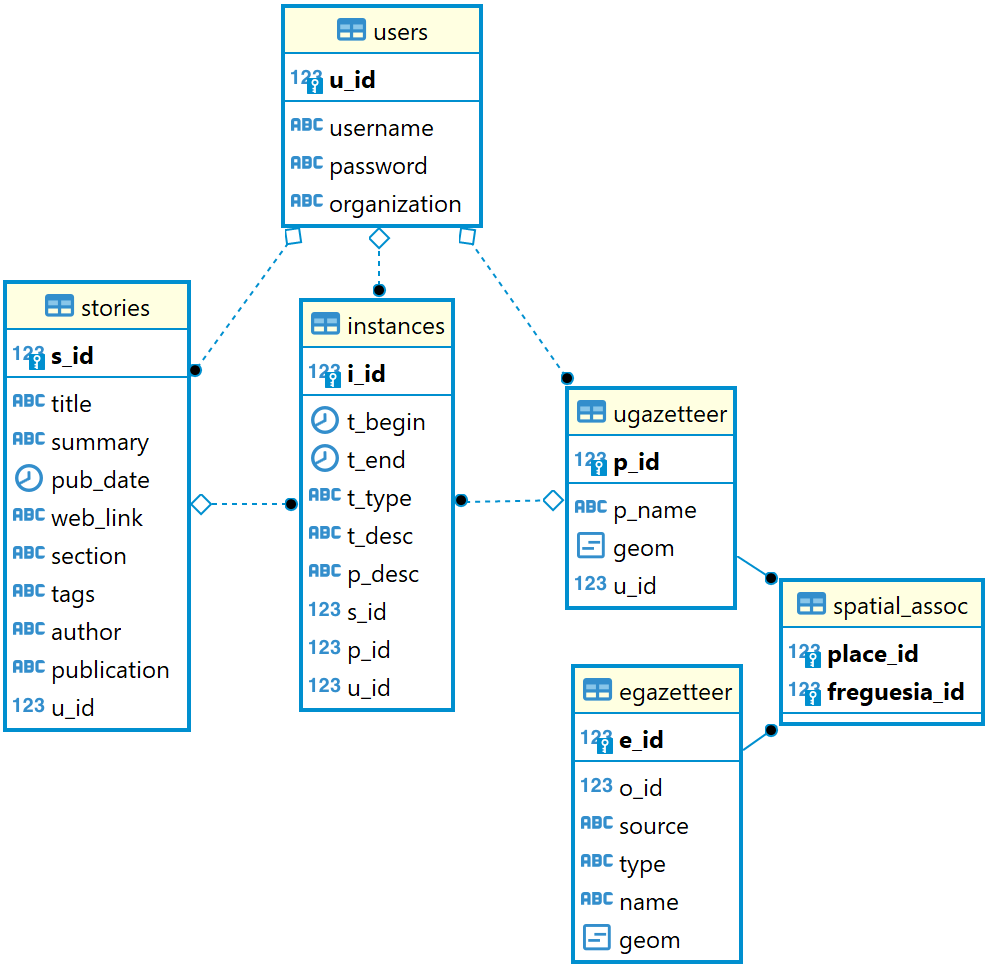

Figure 1: Data model

Publishers can associate spatiotemporal and thematic attributes to the stories they publish, reference to which is stored in a geodatabase. Only publishers perform any sort of data manipulation (create, update, delete) operations, while the remaining views simply read the data stored within.

4.5 Data Collection

4.5.1 Development Corpora

The following corpora is selected as a representation of event reports for the month of October, 2020. It incorporates public organization (Municipality and selected Freguesias) as well as private newsmedia sources to inform the organiational rubric and data model. As historical values, these are used to develop and test the application.

Source	Data	#	Contents' dates	Scale
CML	Boletim 1392	36	Mar - Oct20	address - municipality
	Notícias	42	Oct20	sub-bairro
JFC	Notícias	21	Sep - Dec20	sub-bairo - freguesia
	Newsletter			
JCE	Agenda	4	Jun-Dec20	freguesia
	Notícias	10	Oct20-Jan21	address - freguesia
Público*	ípsilon	83	Oct20-2030	address - inter-municipality
	impar			
	Público			
	Fugas			
	p3			
	(uncategorized)			
Total		**196**	**Mar20-2030**	**address - inter-municipality**

Table 1: Corpora data

**Público data was procurred via the website's search feature, which includes results from all of their products. The filters used were keyword "LISBOA" published during October 2020.*

All attributes are copied directly from the source materials, with the exeption of times and places. These are extracted from the corpora, as these elements may not be explicitys stated ("yesterday" or "near the road" or "on his birthday"). Required fields indicate those that must be extractable from articles to be included in the corpora. Recommended fields are ideally included, and efforts to extract this data will be made if not immediately obvious. If not applicable, these fields may remain blank.

Attribute	**Type**	**Priority**	**Visibility**	**Note**
Title	Text	Required	Yes	
Summary	Text	Recommended	Maybe	Keyword search priority
Section	Text	Recommended	Yes	Thematic filtering
Themes	Text	Recommended	Yes	Thematic filtering
Times	Date	Required	Yes	Temporal filtering
Places	Text list	Required	Yes	Geospatial filtering
Author	Text	Recommended	Yes	
Publisher	Text	Required	Yes	
Publication date	Date	Required	Yes	Default temporal filtering and extraction
Link	Web link	Required	Yes	

Table 2: Corpora attribute collection

4.5.2 Basemaps

Basemaps are include to orient the publisher or reader, improve the browsing user experience, as well as to provide automated spatial enrichment by automatically associating key values (such as freuesia). Additional base features (such as roads, parks, blocks, etc.) are already accommodated into the search tools associated with MapBox, so they have

not been manually incorporated as a part of this work, though this is an opportunity for further enrichment in the future.

Source	Name	Geometry	SRS
DGTerritório	CAOP 2019 - Continental	MultiPolygon	EPSG:3763
MapBox	Roadways, administrative boundaries, built features, natural features, places, terrain	MultiFeature	EPSG:3857

Table 3: Base map data

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 spatial autocorrelation on a very fine level. This, of course, presents a challenge as there is no formal definition of neighborhoods as they are not officially recognized, nor static. Some applications have defined approximations of these (see the differentiation used by Idealista in their spatial searching functionality, figure 2), but these are not publically 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 user gazetteer developed in this application to inform future approximations of neighborhoods in the future. 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 2: Vector approximation of neighborhoods in Campolide

4.6 Preprocessing

The environment was initialized using python within the Flask app framework.

The test corpora of October 2020 was reviewed and thematic, temporal, and spatial extraction. Feature attributes were adjusted to better fit the actual samples of corpora. A selection of these were added as records to a new shapefile in QGIS 3.8 and saved polygon geometries in EPSG:3857. These geometries were staged for testing and validation throughout the development process.

Models representing each table in the postgris database were defined and initialized. Basemap features such as freguesias are loaded into the postgres database, then copied and culled to the greater Lisbon area (inclusive of all DL and AML areas) as individual records into an External Gazetteer (EGazetteer) table. The values associated with various groupings (DL, AML, and municipalities) were unioned into single features and added as new records. All values were transformed into the EPSG:3875 for consistency.

4.7 Views and functionality

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

figure of publishing view

4.7.1 Publishing view

Publishers are invited to login prior to manipulating the database. This is because their userid is a requirement to commiting new story records to the database, so that only this user (or, in the future, their colleagues) can adjust the articles that they publish. This is also to support monitoring of recent activity.

Publishers are then encouraged to associate thematic attributes to their story which will be immediately saved to the database. Publishers are then invited to define one or more instances (temporal and spatial definitions) to the story on a new window pane. Here, publishers may draw one or more polygons associated with a particular time frame, define the time frame, including optional additional context, and save the instance to the database, associated with the original story. Additional instances may be added as appropriate. The stories and their associated instances comprise the GeoCorpora that is available for browsing, searching, and exporting.

Publishers may then navigate back to their profile where they can view all recent stories, and elect to edit any details as necessary.

figure of publishing view figure of recent activity view

4.7.2 Search view

The GeoCorpora is immediately available for browsing and searching. To access this view, users are not required to login. Users may navigate around the base map at will, zooming in and out to explore the locations of stories, and viewing summary details in the tooltips. By clicking on a particular story, the user will arrive at that stories page where they can view all of the details simultaneously, as well as a context map. From here, users may navigate directly to the original article where it is being hosted by the publisher.

To assist in searching, users may define parameters for any of the thematic attributes. They may also define temporal filters to establish a timeframe of interest for either the content (instances) or publication of the articles. Likewise, they may select to search within predefined areas (from the egazetteer) or design their own search area.

The results are displayed in both list and map format to suite the user's preference.

figure of searching view

4.7.3 Context view

The context view is accessible by navigating to the page associated with each story. There, all of the attributes are displayed along with a small context map to situate the story in the city.

figure of context view view

4.8 Validation

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

Future testing of the system may also include a comparison of natural language processing (such as the system described in [21]) using the same corpora for the same period and comparing the results. The system is expected to outperform GDELT and such other large scale, international endeavors, as these do not attempt hyperlocal granularity.

add these references to methodology:

4.9 Web applications

4.9.1 Design

4.9.2 Open Source

4.9.3 Data

Data access

4.9.4 GIS design

4.9.5 Web services

5 Results

6 Analysis

6.1 Future opportunities

6.1.1 Distribution

The WebApp should be made publicly deployed for interaction with and contribution from any users. The product shoudl be licensed as free and open source such taht it can be accessible and leveraged by other individuals and organizations for further development or related projects. Future developments should continue to leverage open source tools, platforms, and data to support this end.

A plugin to relevant backoffices (such as NewStory on WordPress) 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.

6.1.2 Validation

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

6.1.3 Application to GeoIntelligence

-Potential future use of this toolset: GeoINT source. [41]
-Potential future focus/further development for dashboard monitoring OR the system could feed into such systems, providing high confidence georeferencing as it has been manually defined by the author. [39]

6.1.4 Sub-article definitions

-Future research directions: "spatial role labelling... 'the task of identifying and classifying the spatial arguments of spatial expression mentioned in a sentence'... spatial role labelling is key not only in geographic information retrieval but also in domains such as text-to-scene conversion, robot navigation or traffic management systems." [42]
-"We should experiment with how to visualize uncertainty, possible errors and imperfections in our data. And most importantly, we should keep in mind how data can be a powerful tool for all designers, bringing stories to life in a visual way and adding structural meaning to our projects." [20]

6.1.5 Recommended extraction of place

-Avoid the issue of high volume evaluation all together. Those who write the story can apply the location which is the most accurate option. In the absence of the author assigning place, automatic extraction (perhaps such an LSTM model) is an appropriate tool for historical articles. [43]

6.2 Multiple languages

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 user interface, usch as lproejct 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 langauge opptions of English and Portuguese should support the international use and cross investigation of a wider user base.

6.2.1 Incorporation of historical stories

GDELT

6.2.2 Statistical analysis

-Online clustering: "a clustering algorithm for the news domain should group together all news articles that describe the same news event into groups of articles termed story clusters. Broadly, a news event is defined in terms of both story content and story lifetime - articles in the same cluster should share much of the same important keywords, and should have temporally proximate dates of publication." [39]
-Future development: "We will consider ways to use clustering to determine the news provider's geographic scope (i.e. the geographic location of the newspaper), and use it to improve both geotagging and local news coverage." leveraging local knowledge/gazetteers [39]

6.2.3 User experience

Future iterations of the project should include an improved user experience via a more aesthetically pleasing GUI and

6.3 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.

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 me and future efforts may involve collaboration with one or more cities to design a public participation tool. Additional functionality may include additional languages, additional study areas, development of a smart phone application, an option for automatic localization (such as for geo-tagging news stories or

proximal searching), incorporation of historical datasets, incorporation of future events, additional data visualization options, APIs for integration with other applications, white-labeling options for commercial applications, etc.

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

7 Conclusion

7.1 Further devlopments

7.2 Dissemination of research results

References

[1] Mário J. Silva, Bruno Martins, Marcirio Chaves, Ana Paula Afonso, and Nuno Cardoso. Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30:378–399, 2006.

[2] Devanjan Bhattacharya and Marco Painho. 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, 2018.

[3] P.A. Longley, M.F. Goodchild, Maguire D., and D.W. Rhind. *Geographic Information Systems and Science*. John Wiley & Sons, Ltd., 2 edition, 2005.

[4] DMA WGS 84 Development Committee. Department of defense world geodetic system 1984, 9 1991.

[5] Jayna McQueen Baker, Gabriel Huddleston, and Erin Atwood. The map as object: Working beyond bounded realities and mapping for social change. *Educational Research for Social Change*, 8:138–152, 2019.

[6] Hanfa Xing, Jun Chen, and Xiaoguang Zhou. A geoweb-based tagging system for borderlands data acquisition. *ISPRS International Journal of Geo-Information*, 4:1530–1548, 2015.

[7] A. Rajabifard. 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*, page 9, 2009.

[8] Hao Jiang, John Van Genderen, Paolo Mazzetti, Hyeongmo Koo, Min Chen, Hao Jiang, John Van Genderen, Paolo Mazzetti, Hyeongmo Koo, Min Chen, and Hyeongmo Koo. Current status and future directions of geoportals. *International Journal of Digital Earth*, 13:1093–1114, 2020.

[9] Stéphane Roche and Abbas Rajabifard. Sensing places' life to make city smarter. *ACM SIGKDD Interntional Workdshop on Urban Computing (UrbComp 2012)*, 2012.

[10] Sarah Barns. Joining the dots: Platform inermediation and the recombinatory governance of uber's ecosystem, 2020.

[11] Hussein S. Al-Olimat, Krishnaprasad Thirunarayan, Valerie Shalin, and Amit Sheth. Location name extraction from targeted text streams using gazetteer-based statistical language models. pages 1986–1997, 2018.

[12] Greg G. Brown and David V. Pullar. 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, 2012.

[13] Albert Acedo, Tiago Oliveira, Mijail Naranjo-Zolotov, and Marco Painho. Place and city: toward a geography of engagement. *Heliyon*, 5:e02261, 2019.

[14] Massey D. A global sense of place. *Marxism Today*, June:24–29, 1991.

[15] Guoray Cai and Ye Tian. Towards geo-referencing infrastructure for local news. *Proceedings of the 10th Workshop on Geographic Information Retrieval, GIR 2016*, pages 1–10, 2016.

[16] Marco Painho and Isabel Pina. The invisible cities-can ppgis connect citizens to urban policies? *Revista Internacional de Ciencia y Tecnología de la Información Geográfica*, 1:1–4, 2013.

[17] Peter Williams. What, exactly, is a smart city?, 2016.

[18] Nader Afzalan, Thomas W. Sanchez, and Jennifer Evans-Cowley. Creating smarter cities: Considerations for selecting online participatory tools. *Cities*, 67:21–30, 2017.

[19] Tiago H Moreira De Oliveira and Marco Painho. Open geospatial data contribution towards sentiment analysis within the human dimension of smart cities, 2021. ¡br/¿.

[20] Giorgia Lupi. Data humanism, the revolutionary future of data visualization, 2017.

[21] Gilberto Rivera, Rogelio Florencia, Vicente García, Alejandro Ruiz, and J. Patricia Sánchez-Solís. 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, 2020.

[22] Chris Chiappinelli. Democracy: Fueled by pizza, 11 2020.

[23] Jennifer Evans-Cowley and Justin Hollander. The new generation of public participation: Internet-based participation tools. *Planning Practice and Research*, 25:397–408, 2010.

[24] World Economic Forum and ScaleUpNation. Circular trailblazers: Scale-ups leading the way towards a more circular economy, 2021.

[25] Giorgia lupi.

[26] Elijah Meeks. 2019 was the year data visualization hit the mainstream, 12 2019.

[27] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations, 7 1996.

[28] Katy Börner, Andreas Bueckle, and Michael Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *PNAS*, 116:1857–1864, 2 2019.

[29] Elevating quality journalism digital news innovation fund report 2018, 2018.

[30] Mark Monmonier. *How to Lie with Maps*. The University of Chicago Press, 3 edition, 2018.

[31] Yixuan Zhang, Kartik Chanana, and Cody Dunne. Idmvis: Temporal event sequence visualization for type 1 diabetes treatment decision support. *IEEE Transactions on Visualization and Computer Graphics*, 25:512–522, 2019.

[32] Reinout Kleinhans, Maarten Van Ham, and Jennifer Evans-Cowley. 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, 2015.

[33] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. *Proceedings - International Conference on Data Engineering*, pages 201–212, 2010.

[34] Sophie J. Lee, Howard Liu, and Michael D. Ward. Lost in space: Geolocation in event data. *Political Science Research and Methods*, 7:871–888, 2019.

[35] Luke S. Snyder, Morteza Karimzadeh, Ray Chen, and David S. Ebert. City-level geolocation of tweets for real-time visual analytics. *arXiv*, pages 0–3, 2019.

[36] Eurostat. Eurostate regions and cities - overview.

[37] Institudo Nacional de Estatística. Censos 2021 resultados prelíminares, 2021.

[38] Luís Bonixe. As primeiras experiências de radiofusão local em portugal (1977-1984). *Os média no Portugal Contemporâneo*, 19:183–195, 2019.

[39] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: A new view on news. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, pages 144–153, 2008.

[40] Ramiz Sami. Tools i recommend for building geospatial web applications — by ramiz sami — the startup — medium, 10 2019.

[41] Anusuya Datta. Top six geoint trends, 2018.

[42] Morteza Karimzadeh and Alan M. MacEachren. Geoannotator: A collaborative semi-automatic platform for constructing geo-annotated text corpora. *ISPRS International Journal of Geo-Information*, 8, 2019.

[43] Andrew Halterman. Geolocating political events in text. pages 29–39. Association for Computational Linguistics, 2019.

[44] Jerry Low. How to host your own website: Complete beginner guide, 9 2020.

[45] Richard Nordquist. Exonym and endonym, 2018.

[46] Morteza Karimzadeh, Scott Pezanowski, Alan M. MacEachren, and Jan O. Wallgrun. Geotxt: a scalable geoparsing system for unstructured text geolocation. *GIS*, 23, 2019.

[47] Sarang Gupta and Kumari Nishu. Mapping local news coverage : Precise location extraction in textual news content using fine-tuned bert based language model. pages 155–162. Association for Computational Linguistics, 2020.

[48] Felix Hamborg, Corinna Breitinger, and Bela Gipp. Giveme5w1h: A universal system for extracting main events from news articles. INRA, 2019.

[49] Richard Ishida and Susan K. Miller. Localization vs. internationalization, 2005.

[50] Diogo Queiroz de Andrade and Catarina Carvalho. Interview, 2020.

Appendix

A Relevant terms

Abstract place (AP): A point in space or area non-conforming to current or historical ABs or recognized POIs.

Administrative boundary (AB): A geographical area limit managed by an entity; ex: the municipality of Lisbon, Portugal or the 2nd congressional district in Colorado .

Aliasing: "multiple names refer to the same geographic location, such as 'Los Angeles' and 'LA'." [39]

Attribute: an informative element of data stored in a data field.

- **Spatial attribute (SA)**: a description relating to location; ex: 'where did something happen' or 'where was it logged'.
- **Temporal attribute (TA)**: a description of when; ex: 'at what time did it happen' or 'which day was it published'.
- **Thematic attribute (ThA)**: a description of what, why, or how; ex: 'what happened' or 'who published it').

Ambiguity

- **Referent ambiguity**: "the same location can have more than one name" [1]
- **Referent class ambiguity**: "tthe same name can be used for locations as well as for other class of entities, like persons or company names" [1]
- **Referent ambiguity**: "the same name can be used for more than one location" [1]

Civic engagement (CE): define! [13]

Comma separated value (CSV): 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.

Content management system (CMS):define [44]

Contributor: define!

Data model: a graphical representation of the data structure and relationships definitions.

Data visualization literacy (DVL): define [28]

Endonym: "a locally used toponym" [45]

Even-occurring: "all locations where events occurred regardless of whether the event is the event of interest." [34]

Even-relevant: "all locations where events occurred regardless of whether the event is the event of interest." [34]

Exonym: "a externally used toponym" [45]

Gazetteer: A geographical index relating descriptors to location; ex: GeoNames , which related names of places to geographical coordinates.; Gazetteer: "a geographical dictionary, most commonly containing place names and associated properties such as geographic coordinates, type of place, and population, among others." [46]

Geographic information retreival: [46]

Geographic name ambiguity: "a given name might refer to any of several geographic locations." [39]

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 [47]; "Geocoding is the process of parsing places and addresses written in natural language into canonical geocodes, i.e., one or more coordinates referring to a point or area on earth." [48]

Geographic information system (GIS): A framework for the manipulation and analysis of geographic data.

Geoinformatics: define

Geoparsing: the recognition of place names in text' [1]; Geoparsing: "to enable the use of unstructured text in GIS, place references mentioned in text must be automatically recognized and resolved to the geographic coordinates of those places." [46]; "the process of recognizing place names in text ('toponym recognition') and resolving them to their coordinates or gazetteer entry ('toponym resolution')" [43]; "the process of automatically resolving place reference in natural language (unstructured text) to toponyms in a geographic gazetteer with geographic coordinates. Geoparsing enables the extraction of textual information about places for use in geographic information systems (GIS) and other applications. [42]

Geoportal: A user (usually a jornalist for commercial publications, a government official for public orgnaizations, or an individual for a blog contribution) that assigns spatial definitions to an incident via the *Input* tool. "A consolidated web-based solution to provide open spatial data sharing and online geo-information management" [8]

Geotagging: "The process of identifying and disambiguating references to geographic locations (i.e., toponyms), known as geotagging, 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)." [33]

Ground truth: hand coded set of actual locations for training and verification [34]

Incident: Defined within the project as any content of a news article that has spatial and temporal dimensions. These can be past, present, future, or related to multiple instances in time. Likewise, each can occur in a single place or in multiple places, as a point in space or as an area (polygon), and be associated with a recognizable place (such as an AB or a POI) or over areas not commonly recognized (an AP).

Internationalization (i18n): "the design and development of a product, application or document content that enables easy localization for target audiences that

vary in culture, region, or language." [49]

Localization (l10n): "the adaptation of a product application or document content to meet the language, cltural and other requirements of a specific target market" [49]

Location: define

Location based services (LBS): define

Named entity recognition (NER): "which is concerned with the identifying entities such as person, location, and organization names." [39]; " 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') " [1]

Natural language processing (NLP): define

NeoGeography: define [16]

Open source (OS): 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.

Online participatory tool (OPT): define [18]

Ontology: "a formal, explicit specification of a shared conceptualization" [6]

Place: define

Place attachment (PA): define [13]

Point: define!

Point of interest (POI): any entity (natural or artificial) with a well-defined location; ex: Praça do Comércio or Garden of the Gods.

Polygon: define!

Proof of concept (POC): functional or demonstrative of the basic project concepts. [13]

Reader's spatial lexicon: "those locations that the reader can identify and place on the map without any evidence" [33]

Really simple syndication (RSS):define

Scope: define

- **Content scope**:"the story content's geography" [39]
- **Provider scope**: "the publisher's geographic location" [39]
- **Serving scope**: "based on the reader's location" [39]

Sense of place (SOP): define

Smart city: define

Social capital (SC): define [13]

Spatial data infrastructure (SDI): define [9]

Tag: content, section, or descriptive designations defined by the media publisher; ex: 'política', 'primeiro-ministro', 'governo' (from Público), or 'coronavirus', 'denver', 'homelessness' (from The Denver Post).

Toponym: a textual reference to geographic location [33]

Type by task taxonomy (TTT): [27]

User interface (UI): the method of interaction between a user and the program.

Volunteered geographic information (VGI): Define [9]

Web application (Web App): a program running on a web server that is accessible via a web browser with internet connectivity.

Wireframe: a design mockup of a website to demonstrate functional logic.

B Data visualization

-"quantitative data can be converted into qualitative data (e.g., one may use thresholds to convert interval data into ordinal data). Ordinal rankings can be converted to yes/no categorical decisions (e.g., to make funding decisions). The reverse is possivel as well [e.g., multidimensional scaling converts ordinal into ratio data]" [28]

-"quantitative data can be converted into qualitative data (e.g., one may use thresholds to convert interval data into ordinal data). Ordinal rankings can be converted to yes/no categorical decisions (e.g., to make funding decisions). The reverse is possivel as well [e.g., multidimensional scaling converts ordinal into ratio data]" [28]

DVL-FW	Bertin	Description
Nominal	Qualitative	"support equality check"
Ordinal	Ordered	"assumes some intrinsic ranking but not at measurable intervals"
Interval	Quantitative/Numerical	"the zero point is arbitrary"
Ratio		"there exists a unique and non arbitrary zero point"

Figure 3: Logical mathematical operations permissible for data per Borner2019's DVL-FW

Data Scales	Logical Math Operations				Measure of Central Tendency	Examples
	= ≠	< >	+ -	x ÷		
Nominal	✓				mode	
Ordinal	✓	✓			median	
Interval	✓	✓	✓		arithmetic mean	0-6 | 7-12 | 13-18
Ratio	✓	✓	✓	✓	geometric mean	0 1 2 3

Fig. 1. Logical mathematical operations permissible, measure of central tendency, and examples for different data scale types.

Figure 4: Logical mathematical operations permissible, measure of central tendency, and examples for different data scale types (Borner2019)

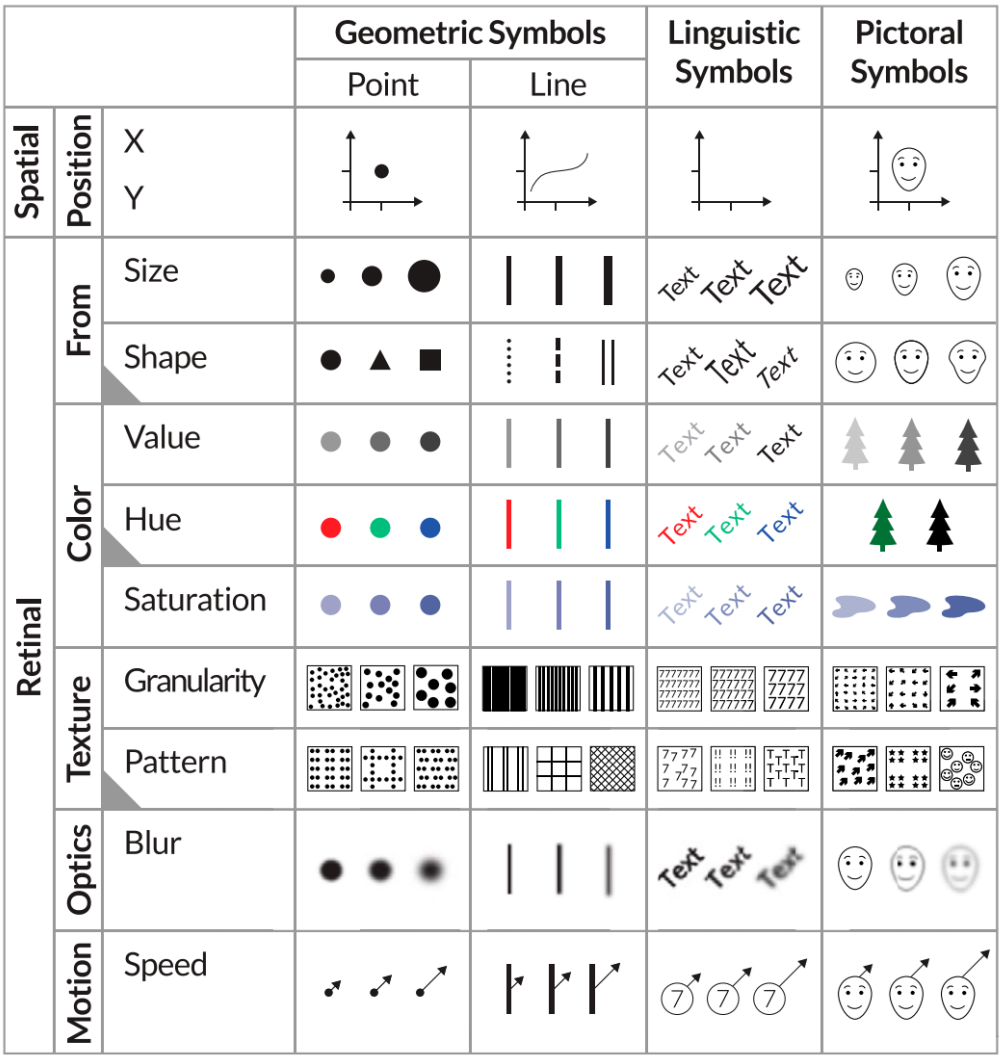

Fig. 3. Four graphic symbols and 11 graphic variables from full 11 graphic symbols by 24 graphic variables set in ref. 34. Qualitative nominal variables (shape, color hue, and pattern) have a gray mark.

Figure 5: Four graphic symbols and 11 graphic variables from full 11 graphic symbols by 24 graphic variables set in ref. 34. Qualitative nominal variables (shapre, color, hue, and pattern) have a gray mark. (Borner2019)

Temporal folding

-"With long, rapidly changing time series it can be difficult to see the subsets relevant to an event of interest. Many visualization excelat revealing periodic or cyclic phenomena through carefully chosen folded time scales, but non-periodic patterns can be obscured by a fixed time scale. When considering multiple event paris across sequences with varied interspersed gaps, it can be difficult to see the overall pattern of relationships between co-occurring events. THese problems are compounded by issues o data quality such as missing data, uncertainty in sensor or manual logs, inconsistency between sources with variou temporal granularities and level of accuracy, adn incorrect timestamps." [31]

-"Temporal folding, or splitting, a sequence into periodic units like hours, weeks, months or years can be used to find cyclic phenomena. Folding can reduce pattern variety facilitating visual analysis." [31]
-"Aligning sequences by sentinel events of interest helps users identify precursor, co-occurring, and aftereffect events... When aligned by a single event we can maintain a consistent time scale between folded or reconfigured units of event sequences. However, it can be valuable to explore the sequence between two separate sentinel events." [31]

C Preliminary specification

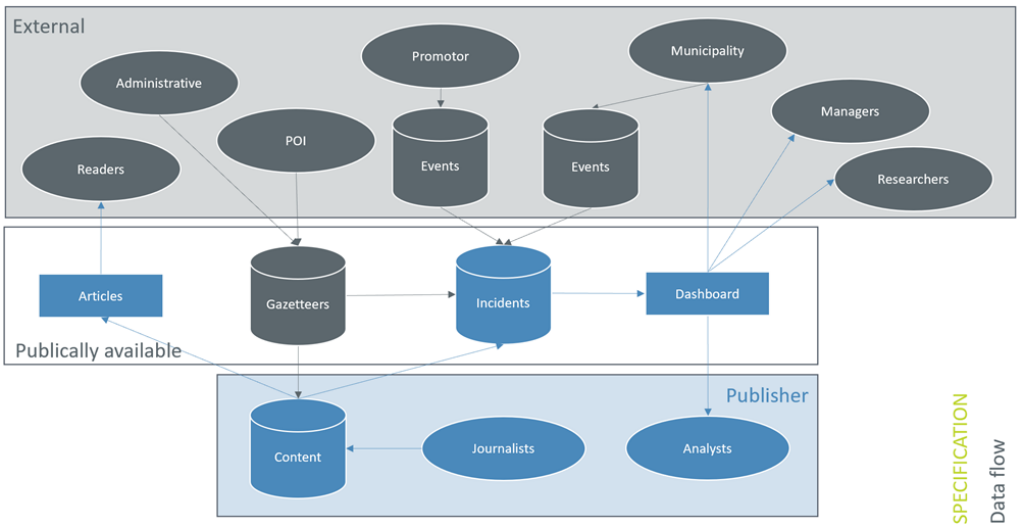

Figure 6: Preliminary data and information flow

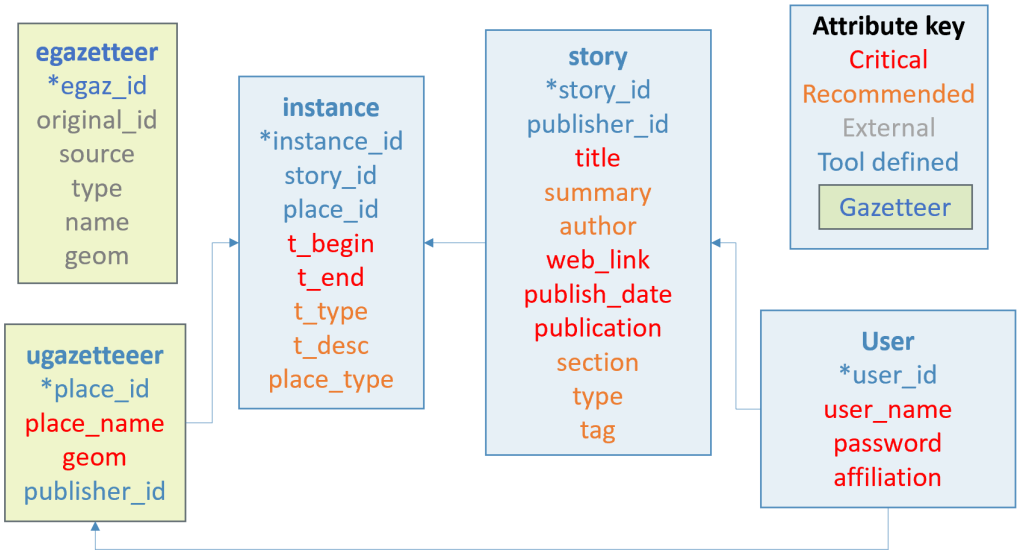

Figure 7: Preliminary data model fo the spatial database

Figure 8: Preliminary *Input* layout

Figure 9: Preliminary *Context* layout

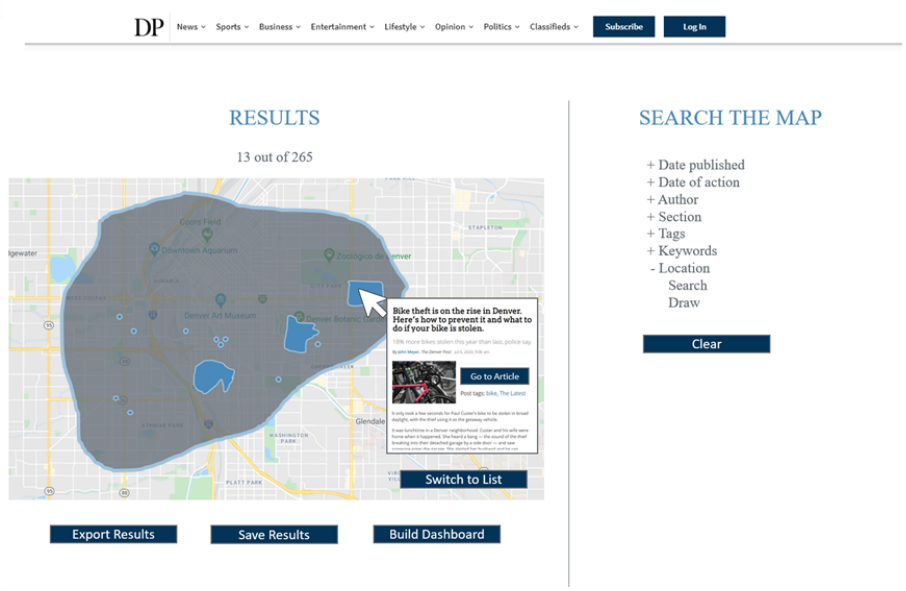

Figure 10: Preliminary *Search* layout

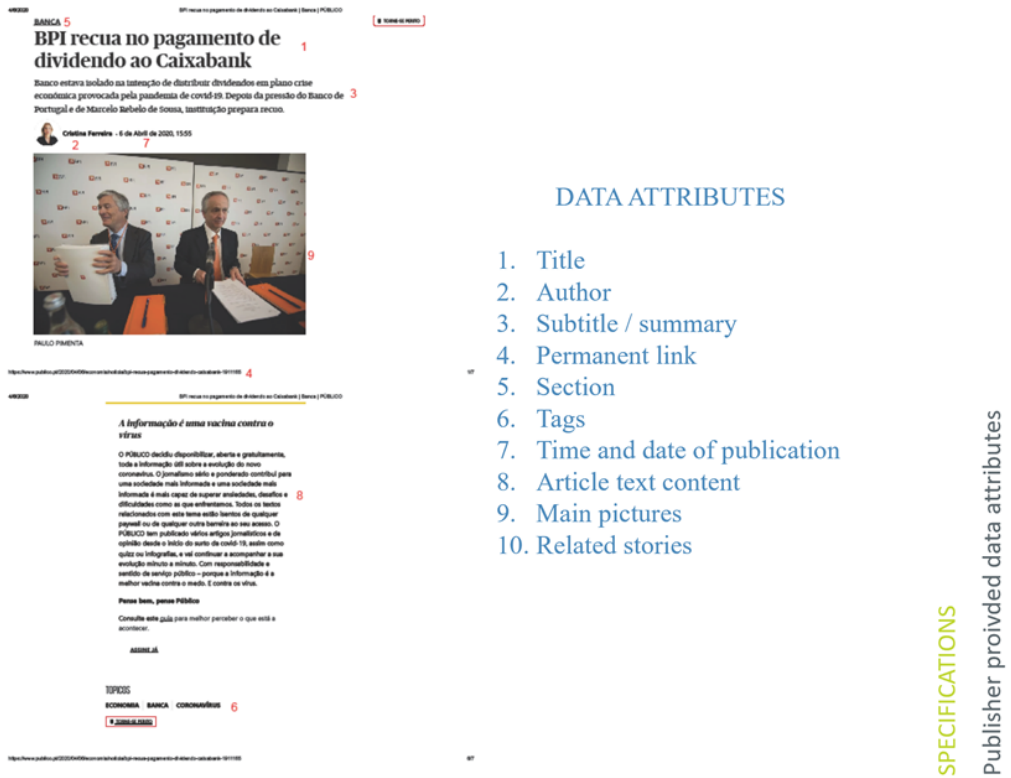

Figure 11: Publisher provided data attributes

D Example GUIs

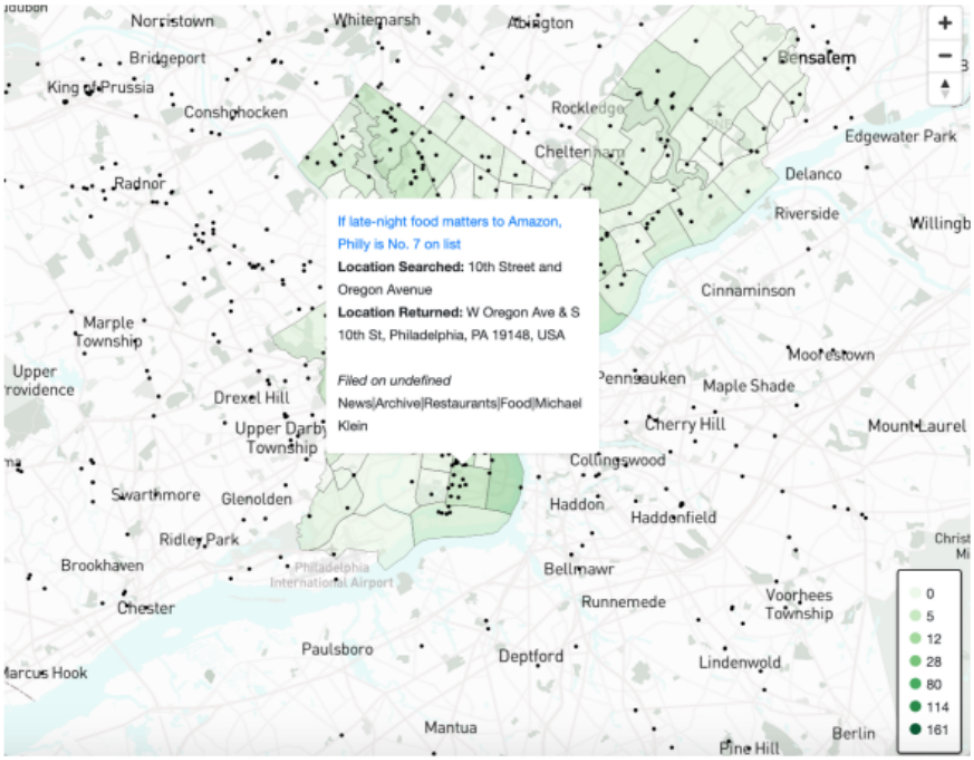

Figure 6: 1000 articles from Philadelphia Inquirer plotted on the map. The green polygons represent Philadelphia neighborhoods with color indicating number of times they have been referenced. Dots represents the locations referenced. The tooltip shows the metadata for the article that references one of the locations.

Figure 12: Mapped article layout (points) from Gupta2020

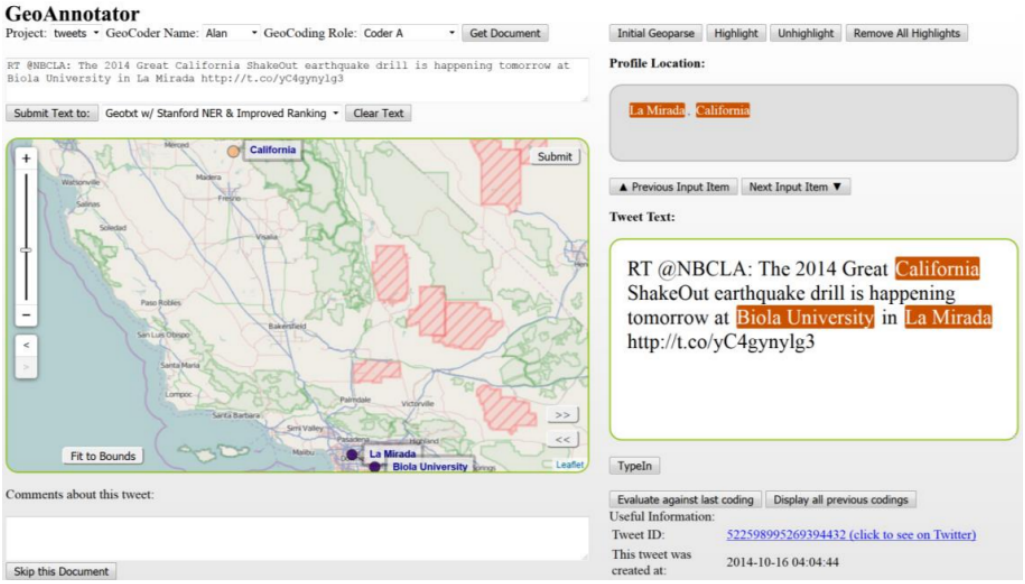

Figure 1. GeoAnnotator's user interface (UI) main window during the initial stages of development. The panel on the right shows the text that is already annotated for named entities and fed into the platform. Place names are highlighted in orange. The panel on the left shows the map interface where highlighted named entities are mapped using a pre-generated list of toponyms through pre-processing. If the default toponym is not the correct one, users can click on it, retrieve a list of "alternative toponyms" and select the correct one. There is often some context that makes it possible for a human annotator to identify the correct toponym, either within the tweet itself, from the profile information and/or from URLs that may be included in the tweet. The interface provides a link to the original tweet on Twitter so that the annotator can search for more context if necessary.

Figure 13: Annotation via GeoAnnotator from Karimzadeh2019

Table 1. Special tags in GeoAnnotator to demarcate special cases of toponyms. For each special case, an example is provided with the place name underscored.

Tag Name	Application	Example	Notes
Uncertain semantics	When it is unclear if a name is in fact a place name, attribute, reference to an organization, or a boundary case with mixed word sense. The uncertain semantics tag enables corpus end users to include, exclude or isolate such cases for different research studies.	The rising violence by Rikers Island correction officers . . .	Rikers Island can be interpreted to refer to the island or the Rikers Island Correctional Center Facility (both of which are "places") or to the prison as an organization, as well as a noun adjunct modifying "correction officers".
Vague boundaries	When the place name refers to an area or region whose boundaries are not clearly agreed upon.	Temperatures in Hudson Valley . . .	Sources indicate that there are differences in opinion on the exact bounds of Hudson Valley.
Not in gazetteer	When the place name in text does not exist in the gazetteer yet, or is so vaguely defined that addition to gazetteer is not justified.	Headed to the West Coast.	Explained in more detail in the following paragraph.
Overlapping ambiguous (always including human annotator assigned surrogates list, enforced by the system)	When human annotators cannot confidently determine which one of multiple candidate toponyms that overlap in space is being referred to. GeoAnnotator allows users to assign multiple toponyms (i.e., surrogate toponyms) to the place name, and apply the "overlapping ambiguous" tag to indicate that these toponyms can interchangeably be used as the resolved toponym for that mention of Lagos (or any other similar situation).	A man just died of Ebola in Lagos.	GeoNames lists three toponyms for "Lagos" in Nigeria: Lagos State (administrative region), Lagos (section of populated place—the city that is within Lagos State) and Lagos Island (within Lagos City, which is within Lagos State). These entities have overlapping geospatial positions and all can be correct assignments.
Non-overlapping ambiguous (with surrogates list)	When human annotators cannot determine which one of multiple candidate toponyms that do not overlap in space is being referred to. Users can assign a surrogate list of potential candidate toponyms to a place name and apply the "non-overlapping ambiguous" tag to indicate that these toponyms can interchangeably be used.	Washington's changing demographics.	Washington may refer to "Washington D.C." or "Washington State", for example. These toponyms do not overlap and it is unclear which one the text author originally meant to refer to.
Non-overlapping ambiguous (without surrogates list)	When human annotators cannot determine which one of numerous candidate toponyms (that do not overlap in space) is being referred to, and there are too many potential candidates to assign as surrogates. Users can select a potential toponym and apply the "non-overlapping ambiguous" tag without providing a surrogates list (making such cases distinguishable to corpus users, who may exclude or use the cases for special studies).	Springfield feels like spring!	Without additional context, Springfield may be referring to numerous toponyms in different geographic regions.

Figure 14: Annotation ambiguities from Karimzadeh2019

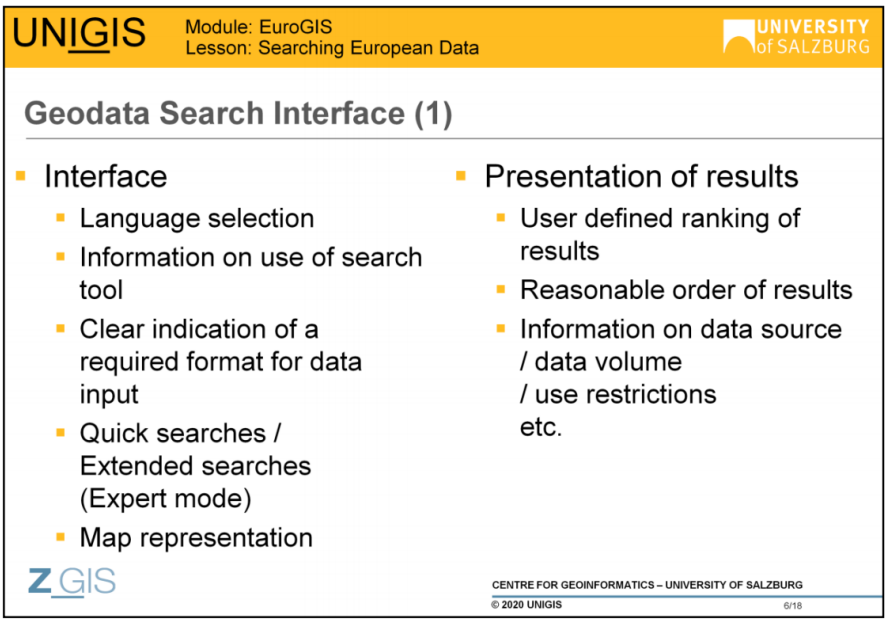

Figure 15: Geodata Search Interface from Eisl2020

E Example Architectures

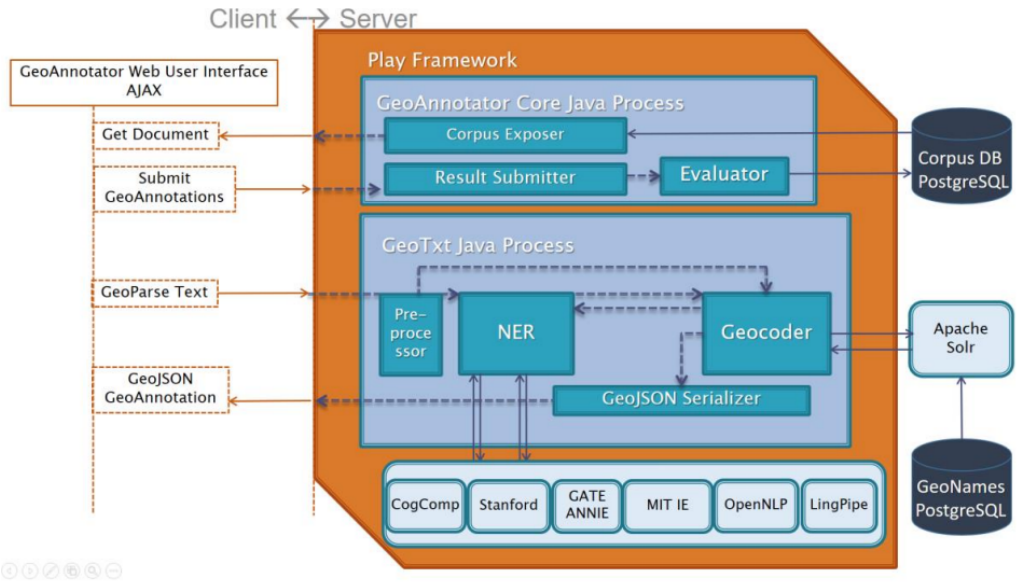

Figure 15. General architecture of GeoAnnotator. The Play framework is used to expose Java application programming interfaces (APIs) as HTTP web-service endpoints.

Figure 16: general architecture of geoannotator from Karimzadeh2019

-Client server model: server implemented in Java, and client (UI) implemented as web-pages (HTML5, CSS and JavaScript). "to be scalable and accessible on a web browser

to many annotators." Bootstrap to be for display size flexibility, leaflet for mapping, jquery/jQuery UI/Rangy for functional operations. Server: PostgreSQL dataabase, corpus exposer module, result submitter module, evaluator module [42]see section 4.2 for architecture description and inspiration

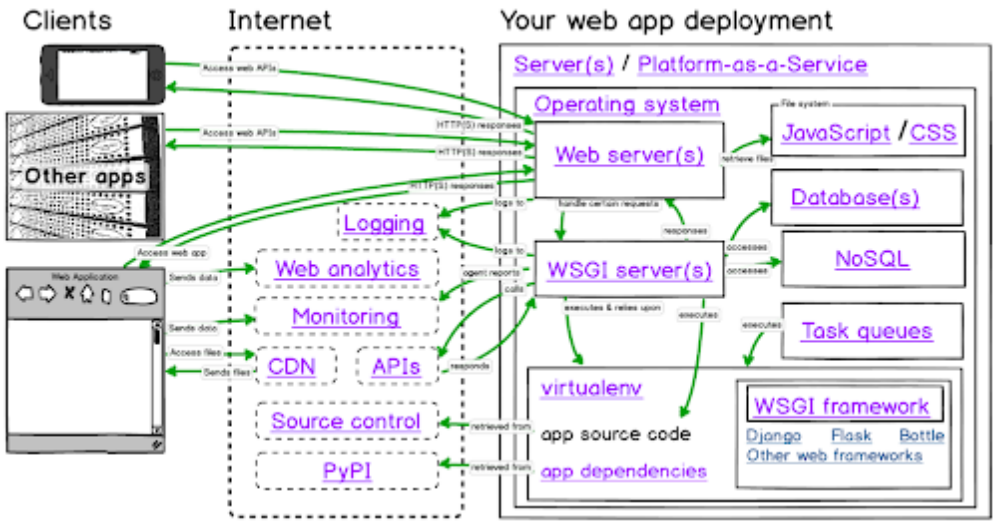

Figure 17: Full stack deployment map from Makai2020

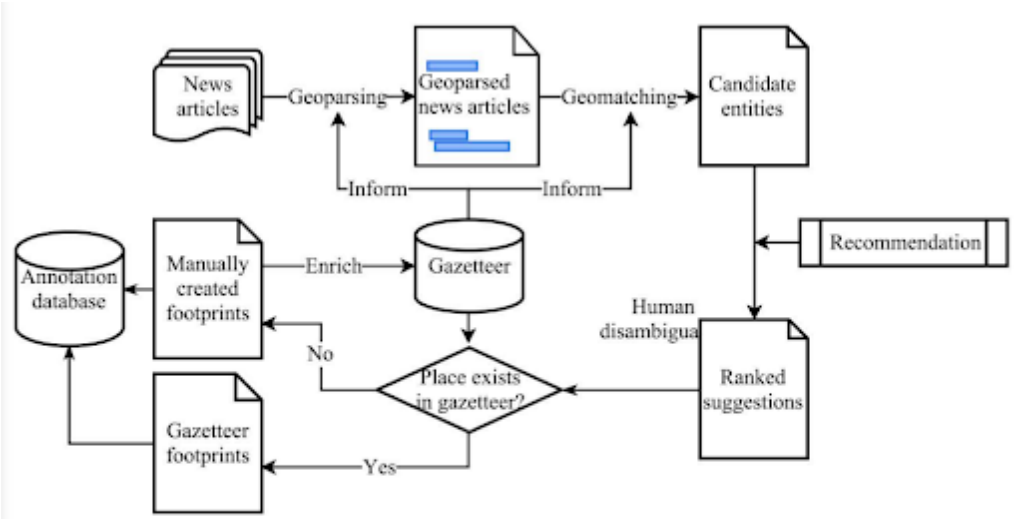

Figure 5: Progressive Geospatial Referencing Framework

Figure 18: Progressive georeferencing frameowrk from Cai2016

F User stories

F.1 Micronews Portal

[50] - Uses Wordpress and NewsPack to publish news stories
- Require a plugin for incorpating main brands of maps (google maps, infogram, mapquest, storymaps, etc.)
- Uses geogrpahical data thatalready exists and is accessed. Want to understand which areas in Lisbon have more traffic crossing, number of stories, super markets, etc.
- Want to pull in elements from Maps.Me
- Have concerns about exposing which areas are covered versus not (news desserts)
- How do you connect with people in different areas: news letters vs. push alerts. Notifying of incidents in areas of interest/proximity. Frustrations with push alert communication.
- Seek to map geographic infomration in the website: a visualization of incidents occurring on a map. All content georefereced in and around Lisbon.
- Have three jouranlists in the field and data from the municipality
- References:

  - functionality of In Your Area to apply to communication of covid cases per area
  - Oriental: orienting people in the middle east

- Lisbon data

  - Access of juntas de freguesias. All data comes from City hall. Covid data come from health minister
  - Usability: many people don't know the covid situation in their neighborhood. Are there 2 or 200 cases? TOOL DEVELOPED INFO APPLICATION STUDY. COVID specific: map cases AND news AND nursing homes AND layered data.
  - Lisboa data: Programa renda acessível