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 Citizen empowerment

2.1.1 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 [2–4]. In fact, “[a]n active and engaged citizen is indeed the main driving force of a ‘smart city”’ [5]. 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 [2]. Beyond efficiency, citizens require safer, more enjoyable living experiences in all aspects of their lives. Governments may accommodate the public interest [4] 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 [5]. 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 [3]. 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 [6].

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” [3] has an internal structure [7], 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” [7], 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” [5].Data in general is already highly regarded as a key comodity for developing an economy [8], 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 [3].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 [9]. 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 [4]. Members of such knoweldge societies [10] , 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” [11] in nontraditional or niche applications [12]. 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.

2.1.2 Communication

In the course of its operations, a smart community should facilitate a ”shared understanding of what is happening” within it [10],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 [13]. 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 [13].

Stemming from the assumption that ”storytelling is the most effective way to merge meaning and emotions” [14], a tremendous and increasingly more ubiquitous tool for effective and relatable communication is data visualization [8, 15]. 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 [16]. Studies indicate that readers prefer pictoral and summarial forms of information (as opposed to purely textual) [13]. Visuals can provide additional context, identify changes, reveal patterns, and showing and distinguishing between relationship types [17], ultimately ”connect[ing] numbers to what they really stand for: knowledge, behaviors, people” [8]. Users, whether they be the general public or decision makers, are expected to have some data visualization literacy (DVL) [18]. 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” [19], 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 [15, 16]. 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 [18]. 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 [18]. They illucidate spatial relationships using layers of data contextualized by basemaps (such as raster images or vector representations) [6, 12]. 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 [17].

In any case, data should be used and interpreted cautiously. Data records are an abstraction of the real world [8]. Often, visualization of data disclude elements of uncertainty [16] or are developed prematurely (without proper analysis) or improperly (misleadingly) [18, 20]. 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 [8, 18, 21].

2.1.3 Public participation

Public participation is a critical element of citizen empowerment. democratic vibrancy, and innovation [4].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 [22]. Further, participation strengthens a community by building social capital amongst its participants, demonstrating trust between members [13]. 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 [4, 13]. It can also prolong interactions, allowing all stakeholders to reevaluate options and motivations throughout the entire process [4]. 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 [22]. 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 [23]. 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 [23]. "[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 [24]. 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" [25].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 [4].

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 [3]. 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, 22]. 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 [22], 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 [24].

2.2 Spatial Components

”Location is involved with everything” [26]. Most data records inherent include some element of place, such as describing where events take place or the monitoring of states at a location. Thoughunderutilized, spatial data plays a key role in decision at the personal, organizational, and regional levels [26]. Geographic information systems (GIS) are computer programs that suppor the collection, sharing, processing and visualiation of geospatial data and its resulting information [12]. This georeferencing (associating data to a map) is fundamental for understanding where people, things, and events are, were, or may be [9, 27]. 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 [6, 12].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 [9, 26]. It is no surprise, then, that location data is already considered valuable and increasingly being incorporated into at all scales of community operations [3, 26]. Spatial information, then, is critical for making educated decisions on key human issues [9].Though this clearly applies to decisionmakers in their respective fields, access to location services is must be a given for all modern society [9]. ”[A] city could not be smart without spatially enabled citizens” [?], who are able to contextualize their own experiences and needs in relation to the realities of their peers.

-“The spaces in which we live, work, and play affect our lives both positively and negatively, and GIS provides a way to visually convey relational meaning of spaces.” [12]

-“Though GIS studies have the potential to represent a fluidity through constant access to technology that written reports often lack, they are still bounded by restrictions of language and symbols, as well as technical and financial resources.” [12]
-“Decades of technology R&D, including, not least, US Government investment in relatively ‘open’ data made available via Global Positioning Satellite (GPS) infrastructure have unleashed the potentials of real-time location tracking and the myriad services this data can be used to enhance, via machine learning techniques.” [28]

2.2.1 Location

-“Location plays a vital part by helping understand relations between datasets.” [26]
-“In context-aware computing, location is a fundamental component that supports a wide-range of applications.” [29]
-“Locality is used to describe a more precise area.” [30]

-Hyperlocality: “a spatiality that is endemic - i.e., locationally specific - to the individual, real-time positionalities of digital platform users.” Ex: “one of the most immediate examples of hyperlocality is the ‘blue dot’ on the mobile Google Maps interface, which visually centers the map data on the real-time location of the user.” [31]

2.2.2 Place

- “From a practical point of view, and in the smart city context, “spatial enablement” refers to the individuals’ (or collective) ability to use any geospatial information and local technology as a means to improve their spatiality, that is to say, the way they interact with space and other individuals on/in/through space. Spatiality is the dynamic component of place making.” -“Place is often uses in the sense of community or neighbourhood, implying an informal relationship to an area surrounding the individual’s place of residence” (refers to Goodchild [8]) [3]
- “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.” [5].
- “Indeed with the exponential growing of location-based social networks (geosocial), Geoweb 2.0 and geoinformation crowdsourcing, citizens are increasingly involved in the the production of geographic information. This kind of information, voluntarily produced and diffused by people, mainly refers to the places they live/use. Indeed, people whom live in a place are often the “experts” of this place.” [3]
- “Spontaneous and localized contributions of individuals (localized tweets, Facebook Places or foursquare chek-in) are most often materialized by Points of Interest (POI). This POI could be considered as new forms fo spatial projections of social relationships and human spatiality.” [3]
- “We find that the sense of place is significant and positively correlated with social capital, while the latter also significantly explains civic engagement at the individual level.” [22]
-“the citizens’ perception of pre-established administrative boundaries can differ from the “real” one nd, consequently, whole administrative boundaries might not cover the SoP, SC, and aCE of all its dwellers.” [22]
-NeoGeogrpahy: Location based services (LBS) and volunteer geographic information (VGI) “This change of paradigm of functioning of the Internet, combined with strong social networks diffusion, changed the concept of community.” [24]
-“How, in the face of all this movement and intermixing, can we retain any sense of a local place and its particularity? An (idealised) notion of an era when places were (supposedly) inhabited by coherent and homogeneous communities is set against the current fragmentation and disruption.” [7]
-Globalization, “speeding up and spreading out” = Time-space compression. “Movement and the communication across space, to the geographical stretching-out of social relations, and to our experience of all of this.” [7]
-“What is it that determines our degrees of mobility, that influences the sense we have of space and place?” [7]
-“A ‘sense of place’, of rootedness, can provide - in this form and on this interpretation - stability and a source of unproblematical identify.” [7]
-“We need, therefore, to think through what might be an adequately progressive sense of place, on which would fit in with the current global-local times and the feelings and relations they give rise to, and whcih would be useful in what rae, after all, political struggles often inevitably based on place. The question is how to hold on to that notion of geographical difference, of uniqueness, even of rootedness if people want that, without it

being reactionary.” [7]
-Places don’t have “single, essential identities”, nor are they based on “internalized origins”. [7]
-“While [a place] may have a character of its own, it is absolutely not a seamless, coherent identity, a single sense of place which everyone shares. It could hardly be less so. People’s routes through the place, their favourite haunts within it, the connections they make (physically, or by phone or post, or in memory and imagination) between here an the rest of the world vary enormously. If it is now recognised that people have multiple identities than the same point can be made in relation to places. Moreover, such multiple identities can either be a source of richness or a source of conflict, or both.” [7]
-“On the one hand communities can exist without being in the same place - from networks of friends with like interests, to major religious, ethnic or political communities. On the other hand, the instances of places housing single ‘communities’ in teh sense of coherent social groups are probably - and, I would argue, have for long been - quite rare. Moreover, even where they do exist this in no way implies a single sense of place. FOr people occupy different positions within any community.” [7]
-“what gives a place its specificity is not some long internalised history but the fact that it is constructed out of a particular constellation of social relations, meeting and weaving together at a particular locus.” [7]
-“If places can be conceptualised in terms of the social interactions which they tie together, then it is also the case that these interactions themselves are not motionless things, frozen in time. They are processes.” [7]
-“places do not have single, unique ‘identities’; they are full of internal conflicts.” [7]
-“The specificity of place is continually reproduced but it is not a specificity which results from some long, internalised history. THere are a number of sources of this specificity - the uniqueness of place. There is the fact that the wider social relations in which places are set are themselves geographically differentiated. Globalisation (in the economy, or in culture, or in anything else) does not entail simply homogenisation. On the contrary, the globalisation of social relations is yet another source of (the reproduction of) geographical uneven development, and thus of the uniqueness of place.” [7]
-“There is specificity of place which derives from the fact that each place is the focus of a distinct mixture of wider and more local social relations. There is the fact that this very mixture together in one place may produce effects which would not have happened otherwise. And finally, all these relations interact with and take a further element of specificity from the accumulated history of a place, with that history itself imagined as the product of layer upon layer of different sets of linkages, both locand and to the wider world.” [7]
-“It is a sense of place, an understanding of ‘its character’, which can only be constructed by linking that place to places beyond. A progressive sense of place would recognize that, without being threatened by it. What we need... is a global sense of the local, a global sense of place.” [7]
-Psychocraphpics: “the prevailing interests of people in the area” [11]
-“Understanding that meaning shifts through an interplay of social, linguistic, and material interactions, GIS has the unique ability to show fluidity yet connectedness across geographic borders, providing intriguing opportunities as a research tool.” [12]
-“the potential to build understanding of spatial context as both socially and politically constructed, but also the tendency to see a map as bounded data in finality.” [12]
-“Place is constructed meaning within [physical spaces]. These places of meaning are a crossing of ‘discursive, interpretive, livesd and imagined practices. Spaces are made up of places with socially created, accrued meaning. Places support the formation of identity,

security, and belonging within spaces. Connecting with ideas of new materialism, spatial researchers see this formation as fluid rather than fixed... Humans impact and are impacted by the social relations in those spaces.” [12]
-“Looking at the politics of spaces helps policy makers, community members, educators, and students understand oppressive cycles of policies imposed on communities.” [12]
-“spatial theory offers a glimpse into the complicated relationships between humans and their surrounding contexts. GIS could be first, a snapshot of those relationships and second, a conversation starter for further investigation and action.” [12]
-New materialism: “the material world is not lifeless soil but something more dynamic.” [12]
-Maps “are objects of analysis that are linguistically and materially [hypercontexually] entangled... resulting in a continual state of becoming. In other words, the material does not construct our reality any more than the discursive defines us socially – instead, it is how the relationship morphs, depending on the context.” [12]
-“When objects, words, people, or places come together, they make something altogether new.” onto-epistem-ology: “the ability to use maps as meaning-making of marked experiences in time an dspace, with the understanding that this is not a static representation of meaning but a point in an evolving state.” [12]
-GIS is a tool to powerfully represent data and effect change by demonstrating how the spatial affects people, and new materialism forces us to concede GIS visual data as already obsolete. The relational ontologies of new materialism disrupt the static representation of a fixed point on a map, bringing the geography and biology of social life into conversation ‘in a discipline which for a long time intentionally decided to ignore the issues arising from this connection.’ “ [12]
-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.” [32]

2.2.3 GeoIntelligence

-Disaster response: “66% of enterprises rank Location INtelligence as either critical or very important” [33]
-Deriving and delivering intelligence: extraction/collection, integration/fusion, iltering/cleaning, enrichment/analysis, distribution/consumable [33]
-“Geospatial intelligence, or the frequently used term GEOINT, refers to the discipline of exploitation and analysis of satellite imagery and other forms of earth observation data to describe, assess, and visually depict physical features and geographically referenced activities on the earth. The term GEOINT is typically used for the defense and internal security domains and offers the capability of monitoring, predicting and countering threats, while helping strategize and support various field operations.” [34]
-Intentionally georeferenced data can contribute to GEOINT as a geolocated data source in addition to other imageries and info fontes. [34]
-Trends: AI and automation, social media and mobile data, analytics-as-a-service, drive for cloud, short shelf life of technologies, geospatial information science [34]
-“Smartphones in every hand, increasing use of social media, and sensors in transport vehicles have turned human beings into sensors. And this is what is driving the next wave

of strategy and innovation for the GEOINT community. With the convergence of new sensors and social analytics, it is possible for defence organizations to access multi-source data for enhanced decision making. As more terrorist/criminal organizations take to social networks like Twitter and Facebook to communicate their efforts, we have seen defence and intelligence organizations are using the right data analytic tools for mining this data to gain a better operational picture of enemy activity." [34]
-Geospatial Information Science: "Different people behave in different ways and culture and location plays an important role in it. With terrorist/criminal activities rising, it is becoming paramount for defense and intelligence communities to learn how humans impact their environments and vice versa; how often cultural norms differ in relation to various environments. This will also enabler he GEOINT community to effectively use technology to capture, manage and analyze geographic information that supports leaders in making informed decisions in complex environments." [34]
-"Who? What? Why? When? Where? These questions form the basis of human exploration. They are fundamental to knowing and understanding our world. Their answers are essential to information gatherings, storytelling,and problem solving." NGA SHOW THE WAY video [35]
-"Everything on Earth, from its watery depths to its hughes peaks, can be measured in space and time, whats known as Geospatial Intelligence or "GEOINT." [35]
-Situational awareness. "Tell me everything that is happening" prior to forming specific questions [36]
-"Hot spot analysis refers to the use of geostatistics and time series analysis to find out the most significant spatial hot spots of a certain indicator and their respective temporal trends." [36]
-ORACULO peacekeeping leverages remote sensing imagery (IMINT), relationship development (HUMINT), reconnaissance, and open sources (OSINT). OSINT: "Thus, though every other way of collecting events promises more reliable results, if we want fast, better than nothing results, we have to turn to open sources, such as news websites." [36]
-"To whom does the information flow? The Media, of course, which has continued to operate throughout the conflict." [36]
-"Extract events from open sources / manage event geodatabases / mine event space time patterns / improving situational awareness for a given context." [36]
-Merge multiple reportings of the same event based on temporal, spatial, and thematic proximity [36]
-"Without location context, first responders are unable to decide where or how to respond to information they receive. This is especially true during natural disasters when geographic content is necessary for dispatching appropriate emergency response." [37]

2.2.4 Cross border mapping

-"cross-border mapping can contribute to cross-border cooperation." [38]
-"The crucial point of transboundary spatial analysis is data heterogeneity." [38]

2.2.5 Feature types

-Massey 1991: "highlights the problem of recurrently trying to draw boundaries to the conception of place and place-related concepts that, in inherently, distinguishes between an inside (e.g., us) and an outside (e.g., them). She also supports that there is no need to conceptualize boundaries in order to define place, advocating that place is a process of social interactions.” [22]

-Polygons (vs. points): ease of implementation, “better encompass of a high range of spatial scales, (from an armchair to the whole earth)”, “better performance of polygon features when there is a limited spatial dataset.” [22]

-“A particular problem with this conception of place is that it seems to require the drawing of boundaries. Geographers have long been exercised by the problem of defining regions, and this question of ‘definition’ has almost always been reduced to the issue of drawing lines around a place.” [7]

-“Instead then, of thinking of places as areas with boundaries around, they can be imagined as articulated moments in networks of social relations and understandings, but where a large portion of those relations, experiences and understandings are constructed on a far larger scale than what we happen to define for that moment as the place itself, whether that be a street or a region or even a continent. And this in turn allows a sense of place which is extroverted, which includes a consciousness of its links with the wider world, which integrates in a positive way the global and the local.” [7]

-“places do not have to have boundaries in the sense of ivisions which frame simple enclosures... Definition in this sense does not have to be through simple counterposition to the outside; it can come, in part, precisely through the particularity of linkage to that ‘outside’ which is therefore itself part of what constitutes the place.” [7]

-“Borderland is a natural transition and convergence area where people, goods, services, and ideas flow across boundaries or seas from state to state. Such cross-border commonalities, which cannot be divided by politically dictated and artificial boundary lines, potentially contribute to sustainable development within the world.” Definition of borderland to describe areas that don’t convene to administrative boundaries. “A borderland region generally refers to the land area adjoining and outside state boundary lines, or hte ocean area among maritime neighbors, and also has different characteristics or geographic conditions than the inner or central parts of the neighboring nations.” [27]

-“Any static method used to represent reality, whether linguistically, discursively, or culturally construed, distorts reality in some way. Researchers acknowledge the material experiences of participants while anticipating that these realities flow and change.” [12]

-“the same PPGIS attributes identified by points and polygons will converge on a collective spatial ‘truth’ within the stuy area provided there are enough observations, however, the degree of spatial convergence varies by PPGIS attribute type and the quantity of data collected.” [23]

-“The use of points for mapping PPGIS attributes and aggregating areas through density mapping constitutes a conservative approach to spatial inference about place significance, but the data demands for point collection are considerably higher than for polygon features.” [23]

-“the spatial feature chosen for soliciting spatial information may influence both the empirical results and the inferences that can be made.” [23]

-“Precision is a measure of the exactness in placing the PPGIS marker on the map... The precision of market placement on the map depends on a number of variables including marker size and map scale as well as participant characteristics such as visual acuity and physical dexterity. Flexible mapping environments that provide multiple map scales and marker sizes such as Google Maps can, in theory, enhance the precision of marker placement.” [23]

-“For PPGIS attributes used in regional planning applications conducted at a larger scale, concern with mapping precision is relatively small compared to the accuracy of the geo-

graphic area represented by the marker. Accuracy reflects how well the marker approaches the true spatial dimensions of the attribute being mapped. Accuracy in PPGIS is influenced by a number of variables including the nature of the PPGIS attribute being mapped (i.e., clarity in operational definitions and instructions enhance accuracy), the quality of the mapping environment (e.g., what base map features are included), and respondent characteristics such as map literacy. " [23]
-"The nature of the PPGIS attribute being identified, in particular, affects accuracy." What does a point representation of a line or area or other fluid definition mean? "the level of accuracy may be indeterminant". How hard are polygon boundaries? [23]
-"Limiting PPGIS attributes to landscape features where accuracy can be objectively verified is not a solution to the accuracy problem because some of the most relevant PPGIS attributes for land-use planning and management are respondent perceptions whose accuracy is not easily determined." [23]
-"In mapping a PPGIS attribute with a point, the spatial attribute of interest is presumed to extend outward from the point in some unknown distance in some unknown direction. For PPGIS attributes identified as polygons, the participant is required to create boundaries that necessarily bifurcate the PPGIS attribute on the landscape, many of which are best viewed as continuous. For polygons, one can argue that the inaccuracy of a single point to represent the spatial attribute is replaced by the inaccuracy of an infinite number of points along the polygon's edge. Alternatively, one can argue that some areal boundaries are better than none and increase the accuracy, even if indeterminate, of the attribute being identified." [23]
-"Because point-based PPGIS attributes can have highly variable spatial distributions, density heuristics may need to be developed for the specific PPGSI attributes being measured." [23]
-"The simplicity of point placement for the participant in PPGIS necessarily results in greater complexity in spatial interpretation for the analyst." [23]
-"The trend line [agreement] has a poor fit when there is such large variation in the size of the polygons and the potential spatial error (amount of polygon area outside the point area) can be up to 18 times greater than the point hotspot area." [23]
-"The quasi-experimental results presented herein demonstrate that the same PPGIS attributes identified by points and polygons will converge on a collective spatial 'truth' within a region provided there are enough observations in the study area. However, points and polygons do present different data information demands and spatial error trade-offs." [23]
-"The use of points for identifying PPGIS attributes and aggregating areas through density mapping constitutes a conservative approach to spatial inference. Using point densities to create polygon areas results in a smaller probability of accepting spatial areas as collectively significant, when in fact they are not." Higher probability of misattributing importance to non-important place with polygons. [23]
-Requires fewer participants (by an order of magnitude) to draw polygons than points to achieve spatial agreement. "However, it is essential that the sample size be sufficiently large to claim regional representativeness regardless of whether points or polygons are the chosen method." [23]
-"PPGIS participants found the placement of points less ambiguous than identifying polygons and thus participants are more likely to complete the PPGIS mapping activity." [23]

2.2.6 Geoportals/mashups

-"There is a big need for spatially referenced data creation, analysis and management." [26]
-"The launch of Google Earth in 2005, and the availability of Google maps programming interface (API), as well as other initiatives, have transformed the way Internet users relate to geographic information (GI). The transfer of the information-creation processes from the specialized domain of Geographic Information Science to the field of action of the non-experts, the fact that other citizens, in addition to the geographers, cartographers or GI specialists, can create their maps with their own content, is to radically change the domains of interest and application of these mechanisms and to impact the criteria for the collection, analysis, implementation and the standard of truth of the information, with implications for information access, participation, power balance and nature of the data." [24]
-GIS: "the opportunity to open spatial information to all stakeholders (presumably leading to better policy-making) and the idea that spatial analysis and outputs (.e. maps) can persuasively convey ideas." [25]
-"innovative and imaginative geo-visualization interfaces such as Google Maps or Open Street Map - made possible by Web 2.0 technologies – have created low-key opportunities for almost any citizen with an Internet connection to generate and publicize their own maps and geographic information." [25]
-"Organizations have different capacities to use new participatory methods. For example, private organizations or planning consultants may be better equipped than local governments to launch an online tool or collect and analyze the data." [4]
-"Internal and external organizational collaborations influence the successful application of new technologies and methods within institutional systems. Collaboration with outside organizations can facilitate the incorporation of new technologies and data sources." [4]
-"There has been a need to develop automated integral spatial system to sense and categorize events and issue information that reaches users directly."
-"It has been proven that the spatial analysis of data gives more meaning to the information extraction and hence enables easier assimilation of large volumes of data. Presently the systems that implement such processes are limited in effect by not utilizing all the data due to their standalone nature, offline or disconnected design, lack of spatial capabilities, unintegrated approach, and temporally disjoint." [26]
-"The solution could be addressed through integrating: data source, spatial data platform, data understanding, knowledge base, inferencing an dvisuzalition into single, well-connected online real-time system. Such a spatial expert system (ES) with knowledge base (KB) will not only serve the critical research of spatializing developmental works but do so to any research relying on real time data capture and analysis with spatial domain of data being the unique enabler." [26]
-"A new paradigm of more open, user-friendly data access is need to ensure that society can reduce vulnerability to spatial data variability and change, while at the same time exploiting opportunities that will occur." [26]
-"Although research scientists have been the main users of these data, an increasing number of resource managers need and are seeking access to spatial data to inform their decisions, just as a growing range of policy-makers rely on spatial data to develop spatial change strategies. With this gravity comes the responsibility to curate spatial data and share it more freely, usefully, and readily than ever before." [26]
-"The interactive graphical user interface (GUI) allows for data visualization manipulation and sharing." [26]
-THe ability to harness and render this information in a location context is a major chal-

lenge." [26]
-"GeoWiki, a geographical semantic wiki system that was introduced in [10], can parse and store multi-source geographical knowledge, mash up with Google Maps and serve geospatial decision-making." [27]
-"A mashup is a Web application that aggregates multiple services to achieve a new purpose. It is an important feature of Web 2.0 and can be used with software provided as a service. With various mashup techniques, it is convenient for developers to obtain data from a variety of data sources on the web and to integrate these data to build new applications." [27]
-"A map mashup therefore combines at least one map data source or services with added information, often geo-referenced to the map data, to create a new map." [27]
-"Geoportals are a consolidated web-based solution to provide open spatial data sharing and online geo-information management." [6]
-"geoportals usually provide access to distributed data systems, offering maps, data discovery, and data downloads. Some of them are also capable of offering online analysis and processing service, enhanced semantic search engines, and dynamic visualization tools." [6]
-There is an ongoing challenge to effectively manage and communicate the vast and growing amounts of spatial data and geo-information. [6]
-"oVer the past few decades, the concept of geoportals has emerged as one of the key solutions for spatial data and geo-information accessing and sharing." most recently through the internet. [6]
-Geoportal: "a point of access to spatial data and geo-information. It is able to provide a geospatial data inventory linking to an inclusive collection of spatial data, geographic information, online services, and data processing tools. In this article, the use of the term 'geoprotal' refers to the human-to-machine interface performing as a single point-of-access to spatial data and geo-information systems, offering sharing capabilities and connecting between geospatial data providers and end users. It is typically employed as a web-based graphical user interface (GUI) equipped with functionalities for accessing Earth observation data and geographical information." [6]
-Drivers of geoportal advances: "(i) scientific geospatial projects and applications, (ii) international organizations, (ii) governmental agencies, and (iv) commercial purposes". [6]
- International organizations: Sharing data from "heterogeneous data sources". [6]
-Governmental drivers: open government policies. "These portals, which were built by answering the open government call (or named the e-government), act as the gateways, anchors, or major starting sites for governmental data, no matter whether the data is spatial or non-spatial in nature." "Openness and transparency are fundamental to ensuring citizens' trust in their governments. Thus the objective of a government geoportal is to foster greater transparency and accountability, providing information available to the public from digital technologies." Example of Geospatial OneStop (GOS) in the USA. [6]
-Commercial geoportal: connect users to digital resources, including geospatial datasets. [6]
-"Data is the content that geoportals provide to end users. A single geoportal may provide heterogeneous geospatial datasets coming from multiple data sources." Geographical and earth observation dataset types, additional products, models, and aggregation, data harmonization [6]
-"The geoportals can facilitate researchers, government officers, and ordinary users in helping them to find the data they needed, with basic searching services equipped in the geoportal." [6]
-"a geoportal is normally a general-purpose tool to discover spatial data and geo-information,

but often there are communty-specific needs that require customized geoportals with dedicated tools for scientists, policy-makers, and the general public." [6]
-"GEoportals should also play a role as the interface of so-called social spatial data infrastructure. New system architectures (e.g. Linked data and Semantic Web) to establish a shared information space relying on URLs and Resource Description Framework (RDF) can also provide another solution for accessing geoportal data." [6]
-"Geoportal is considered as the entry point and the human-to-machine interface of the data and information management system." [6]
-"Promote a common way of accessing data for an enterprise and any communities that need to make use of that data" [33]

2.2.7 Commercial geospatial platforms

-The definition of the search criteria by the user is usually done based on a form, defining the spatial, temporal and thematic coverage of the searched data. In the context of geodata search the use of maps on which searched areas can be defined have become standard – here the region of interest is selected by identifying a point, a rectangle, a polygon or some pre-defined region (e.g. a country). In most cases alternatively geographical names or coordinates can be searched. Especially the latter approach is less intuitive, but also the use of geonames requires some care as very often the way in which names are written differs depending on the language and special national character sets are used. [39]
-Platform ecosystem: "technologies, platforms, data points, user interfaces, relational infrastructure and application programming interfaces (APIs), service providers, investor-financiers and customers" [28]
-"'the platform' is not only a particular computational structure but also a discursive innovation, which works on behalf of its business owners to recast conditions of value-sharing between companies and their marketplace as ostensibly 'open'. These ecosystem-relationships make possible conditions of algorithmic control, subtly shaping, nuding and 'steering' urban behaviors from infinitesimal to global scales, in ways that point to new asymmetries of urban governance in a world of connected devices, things, people and infrastructures." [28]
-GPS enabled smart phones enables the superior experience of ridesharing to traditional taxis. [28]
-"No local knowledge is required: the app provides the requisite information to fulfill the transaction". [28]
-"the value harvested by platforms is in the data generated through digitally mediated transactions and relationships, which informs the terms dictated by platform companies for wider data-driven applications." [28]
-"The computational architecture put in place across a platform ecosystem hinges fundamentally on the role of the Application Programing Interface, or API, which enables different kind of software to connect." [28]
-"Platform architecture can thus be read as highly recombinatory, ingesting and harvesting data from a highly porous, intermediated platform ecosystem of multisided relationalities and transactions, while underpinned by the protocols and modes of value-exploitation that are closely controlled by platform owners." [28]
-"The tactics adopted by platforms firms to entice users to choose one service over another are described as 'steering tactics". For economists, steering tactics are central to the operations of successful platforms, which devote significant attention towards and investment in strategies that can enhance the overall value of the platform to its users." [28]

-“geolocation organized five ‘platform affects’ for digital platform ecosystems. (i) affective space-times of hyperlocality and real-timeness; (ii) experiential affects of smoothness; (iii) affects of connectedness; (iv) affects of trust; and (v) affective value.“ [31]

-“THe merging of digital location with web content, services, and interfaces is pervasive to the extent that digital location - or simply ‘geolocation’ - now features as a native design, logistical, and organizational component of digital platform interfaces, utilities, and affordances.” [31]

-Social media updates (geotag), ridehailing (geologistics), dating (search radius definition), accommodation (interactive map-based filtering) [31]

-“Initial world by geographers has begun to preliminarily attend to the factors atively underwriting the integration of geolocation with digital economies. Yes this work has been largely market-oriented, engaging geolocation as first and foremost an economic commodity.” [31]

-Platform affect: “informs an understanding of digital platform ecosystems as affectively oriented and as expressing a capacity for orienting affects. Platforms generate, accumulate, and cycle affects within ecosystems; they also draw on the affective capacities of other techno-material formations, such as geolocation, to organize affects for digital ecosystems.” [31]

-“as geolocation is put into circulation, it expresses a capacity to affect - perturb, animate, align, mobilize, organize, dis/assemble - other things, both humans and other-than-human alike. Affect is useful for grappling with what geolocation ‘does’, how it ‘does’ what it does, and to what ends.” [31]

-“geolocation organizes affects for platforms by investing users within platform ecosystems through individual and collective attunements towards using, participating within, contributing to, remaining within, and returning to these ecosystems.” [31]

-“The integration of geolocation in the form of positional awareness as a native component of digital platform design invests individuals in highly structured yet simultaneously ontogenetic hyperlocal microgeographies that solipsistically spatially and temporally centre the user.” [31]

-“These space-times invest users within platform ecosystems by literally positioning them at the centre of their own platform-mediated universe, affectively attuning users towards returning to the platform to regain access to a service brokered solipsistically to their hyperlocalities and real-times.” [31]

-“Certainly, the instantaneousness and hyper-locality of platform mediation is significant to the competitiveness of platforms such as Uber in increasingly crowded platform economy marketplaces.” [31]

-“beyond understanding these space-times in terms of the business strategies that they may underwrite affect theory also capture the ways in which once users are affectively invested in these ecosystems via geolocation-organized space-times of the ridehailing platform, these platform-mediated affects ‘turn back’ on themselves… organizing an affective frame of reference which conditions user’s expectations of future experiences of the platform: a potential ride available frome very possible location within minutes of being requested.” [31]

-Amazon prime: “the ability of platform enterprises to operationalize these space-times for consumers through drone delivery is contingent on a complex of geolocation affordances, data, hardware components, and software technologies operationalized in concert,... also expands the range and possibility of modalities through which Amazon may organize affect that retain consumers within its ecosystem… through the leveraging of geolocation to organized the desired hyperlocal rea-times that align both online and offline consumption

experiences with the expectations of consumers.” [31]
-Geolocation as “an affective capacity as an instrumentality that organizes and aligns experiences of platform ecosystems. This especially pertains to consumptive experiences (of goods and services), which are affectively smoothed through the leveraging of geolocation as a coordinating force that produces seamless digital consumption experiences not only within but also across platform ecosystems.” [31]
-“Streamlining the retail consumer experience across all stages of a purchase, from shopping online through to last mile fulfillment... here, geolocation is leveraged as a real-time coordinating force that links and streamlines multiple stages of an online-to-offline retail transaction so as to render the consumption process as experientially seamless as possible for the end consumer.” [31]
-“Geolocation is also leveraged to organize affects of smoothness across platform an application ecosystems through practices of digital aggregation.” [31]
-“By integrating multiple urban transportation asset providers, infrastructures, and operators within a single app ecosystem, Transit smooths urban mobilities by investing users within a singular, cohesive, integrated digital platform experience that mitigates the impracticalities of switching between multiple branded apps of mobility actos both across transportation modalities... and within a modality class.” [31]
-“Inarguably, digital location has become central to quotidian experiences of digital platforms, be it through the ways in which content is accessed an interacted with (location-aware search, spatial interfaces); the ways in which we generate and contribute content (geotagging, personal location data mining); the ways in which we rely on location-aware devices such as sensors and digital assistant to produce ‘smart’ environments for us (smart cities, smart homes); and also the ways in which we interact with the world (augmented and virtual realities)... DIgital location is also an affective instrumentality that organizes how platforms and platform ecosystems are encountered and experienced by end users, and where thesmoothness of these experiences - as underwritten and facilitated by geolocation - is central to the ways in which users themselves become affectively invested in platform ecosystems, coming to see platforms as indispensable to how they get around cities. “ [31]
-“geolocation also functions as a techno-discursive instrumentality that underwrites the affective expression and enactment of social, familial, and romantic relations of connectedness performed through the sharing and monitoring of personal locations via digital platforms and their affordances of connectivity.” [31]
-“digital location is itself also being leveraged as an instrumentality of securing trust in platform ecosystems.” [31]
-“Mobilizing an understanding of location as implicated in arranging transparency and security form platforms informs a nuanced understanding of how it is that users become invested in platform ecosystems despite the ways in which platforms themselves undermine user confidence in digital systems through practices such as psychometric profiling and microtargeting..., covert user experimentation..., and regular data security breaches, including those involving the theft of personal location data”. [31]
-“As an affective capacity to invest users in digital ecosystems, geolocation is leveraged to build trust in platforms by organizing technics and discourses of authorization (digital transactions), and of securitization (against misinformation).” [31]
-Geolocation “invests developers, marketers, and entrepreneurs in the speculative potential of geolocation - in the form of embedded interactive map objects and interfaces - to organize annexation economies for end users. These attention economies are predicated on investing users themselves in digital platforms by ‘attracting eyeblalls’, expanding opportunities for user engagement by providing additional interactive elements (interactive

map objects), and extending the duration of user visits (how long users remain within a digital ecosystem)." [31]
-"geofilters are a brand growth strategy realized mobilizing real-world location... Here again digital location is being discursively mobilized to affectively attune digital developers and advertisers to the capacities of geolocation to both organize spatio-temporally contingent user attention economies and to ultimately convert that attention into both online and offline consumption." [31]
-"Geolocation expresses a capacity to align and attune corporealities and inorganic materialities in ways that work to invest users in platform ecosystems by organizing and orienting affects for platforms." [31]
-"ecosystem refers broadly to interconnected stakeholders and organizations exchanging knowledge and value, each with their own role in the greater whole. No one, central party controls the ecosystem, but rather, it is mutually nurtured and supported for collective health." [14]

- Uber
- AirBnB

2.2.8 Location extraction from text

-IDentify place name: Stanford Core NLP, OpenNLP; Identify placeal nd disambiguate: CLAVIN, Yahoo placemaker; Geography3 [40]
-"Extracting location names from informal and unstructured social media data requires the identification of referent boundaries and partitioning compound names. Variability, particularly systematic variability in location names challenges the identification task." [29]
-Location name contraction problem: "pragmatic influences on writing style shorten names to reduce redundant content in social media." [29]
-Nameheads: "complex phenomenon of alternate name forms". "1) Appellation formation, 2) explicitly metonomy, 3) category ellipsis, and 4) location ellipsis)." Appellation formation: removing the designator of a specific thing (O PONTE vs. PONTE DE 25 DE ABRIL). Explicit Metonomy removes the thing and leaves only the designator (NOVA vs. UNIVERSIDADE DE NOVA). "Both Appellation formation and metonomy pose disambiguation problems, and require context such as the author's location to resolve." "Category Ellipsis and Location Ellipsis pose delimitation problems that can be resolved with a statistical language model." Category Ellipsis: "the author strips words related to the location category" (LISBOA vs. CIDADE DE LISBOA). "Location Ellipsis occurs when an author drops the specific location reference in the location name". Entity delimitation: "to identify the boundaries of a location mention in the text". Previous efforts invoke the application of heuristics (syntatic and semantic) with fallible results, challenge linking to gazetteers. Collocations: "a sequence of ordered words... neither strictly compositional nor always atomic." "Given a region-specific gazetteer, which retains the same location-context as the text, we can construct a statistical model of the token sequences it contains." [29]
-"Tokenize all location names in the gazetteer to construct the n-gram model and then save the resulting lists of unigrams, bigrams, and trigrams." From the gazetteers, try to predict all the different ways one might reference location (assumed shortening of the official names) and create a new list. [29]
-"The field of geolocation extraction collectively involves many different tasks and analyses to be performed over text. The three main tasks among these are: (i) Location

named entity extraction; (ii) Location named entity resolution; (iii) Event's location extraction." [30]

-Even though several geoparsers such as Cliff-Clavin, Mordecai, and Stanford-CoreNLP have been developed to automatically extract named locations from unstructured English text, location extraction from a text is still a challenging task due to the complexity, diversity, and ambiguity of location information in different languages. However, these tools cannot extract the focus location with good accuracy, and most of them cannot differentiate between different locations in the text - i.e. focus locality versus non-focus locality - and are not language agnostic." [30]

-"While traditional named entity taggers are able to extract geo-political entity and certain non geo-political entities, they cannot recognize precise location mentions such as addresses, streets and intersections that are required to accurately map the news article." [41]

-Named Entity Recognition (NER) taggers: "contain tags to identify organizations, geo-political entities (GPE) and certain non-CPE locations such as mountain ranges and bodies of water from text"; "they are not able to extract precise locations mentions such as addresses, streets or intersections in their entirety." [41]

-"Tools trained on conventional NER models. . . . Have been successful in identifying common named entities. However, challenge comes when high level of granularity is of interest in extracting location entities such as specific addresses, streets, or intersections." [41]

-"Fine-tuning pre-trained language model fo domain-specific machine learning tasks has become increasingly convenient and effective". [41]

-Process: "named-entity tagger for the task of precise location extraction involves fine-tuning an existing neural network on a target dataset. [41]

-"Geocoding is the process of taking input text, such as an address or the name of a place, and returning a latitude/longitude location on the Earth's surface for that place." [41]

-Geotagging: 1. Entity feature vector extraction, 2. Gazetteer record assignment, 3. Geographic name disambiguation/toponym resolution, 4. Geographic focus determination [42]

-"Unlike the spatial information used in a Geographic Information System (GIS), spatial information obtained from web documents is often incomplete and fuzzy in nature. A GIS user can formulate data retrieval queries specifying complex spatial restrictions, while a search engine targets a wide variety of suers who only provide simple queries. However, these users are also in need of retrieval mechanisms for queries with geo-spatial relationships. In order to support this, a first step concerns with assigning geographical scopes to web resources, so that the same resources can latter be retrieved according to geographical criteria." [1]

-"Machine learning provides effective techniques for text classification, involving the automatic generation of classifiers from manually annotated training data. However, with very few exceptions, most work in automated classification has ignored the presence of hierarchically structured classes and/or features." [1]

-"Much of the contextual information that could be used to disambiguate the geographical scopes in natural language texts is absent or external to the texts. The amount of training data per feature is also so low that there are no repeatable phenomena to base probabilistic methods on. For instance, the frequency of location names is in itself not sufficient for a good classification, as the same location name will usually not be repeated, even if the name is important." [1]

-"Disambiguating geographical references in the text and assigning documents with a corresponding geographical scope are two crucial steps in building a geographical retrieval tool." [1]

-“Previous studies have demonstrated that recognising geographical place names in text (usually called geo-parsing) is a crucial precondition for geo-referencing web documents. In language processing, 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’) is referred to as Named Entity Recognition (NER).” [1]

-“Ambiguity is the main problem associated with geographical references in text... ambiguity in geographical references is bi-directiona, as the same name can be used for more than one location (referent ambiguity), and the same location can have more than one name (reference ambiguity). The former has another twist: the same name can be used for locations as well as for other class of entities, like persons or company names (referent class ambiguity).” [1]

-“To be useful, NER systems focusing on geographical concepts should handle the complex issues related to how people use geographical references. Place names lack precision in their meaning, and often vary with time, from person to person, and with the context in which they are used. Many times place names are simply temporary conventions, and people’s vernacular geography if also often vague, as they may also be interested in the vicinity of a place without knowing its exact name. Not only spelling variation are common on geographical names, but also the places those names reference change in shape and size.” [1]

-“About 60% of all data (textual and otherwise) are geospatially referenced”. [43]

-Geographic information retrieval: GIR [43]

-OVerall, soem recurrent issues persist in geoparsing research: methods are not evaluated or evaluated on non-public datasets or using proprietary systems (inaccessible, or behind paywalls): demonstration and evaluation rely on small tasks, or small gazetteers, or small geographic scope: and/or focused harvested corpora are used in evaluations that greatly simplify the toponym resolution task.” [43]

-In unstructured, textual data: “the spatial data is specified using text (called toponyms) rather than geometry, which means that there is some ambiguity involved.” [44]

-“The ambiguity has an advantage in that from a geometric standpoint, the textual specification captures both the point and spatial extent interpretation of the data (analogous to a polymorphic type in parameter transmission which serves as the cornerstone of inheritance in object-oriented programming languages). On the other hand, the disadvantage is that we are not always sure which of many instances of geographic locations with the same name is meant” [44]

-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). Geotagging is difficult because the first step involves understanding natural language, while the second step requires a good understanding of the document’s content to make an informed decision as to which of the many possible locations is being referenced.” [44]

-Toponym recognition: “the most common strategy is simply to find phrases in the document that exist in a gazetteer, or database of geographic locations, and many researchers have used this as their primary strategy.” [44]

-Section 2 for shortcomings of current geotagging strategies [44]

-Fuzzy geotagging: “does not fully resolve toponyms in a single article, instead returning sets of possible interpretations for ambiguous toponyms.” [44]

-“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.” [45]
-“Geographic description in texts reflect human conceptualization and experiences of space and places. Different from other forms of geographical data, text-based spatial description are subject to all sorts of ambiguities that prevent effective use.” [32]
-“Geospatial referencing textual documents refers to the task of discovering location phrases and creating unambiguous representation (or footprints) of the meaning of those textual references.” [32]
-“To create unbiased corpora that capture the complexity of natural language and place name ambiguity and annotate with ground-truth toponyms from high coverage and detailed gazetteers, documents used for testing and training should be manually ‘geo-annotated’ by human annotators, i.e., place names should be recognized (segmented) and manually resolved to toponyms in gazetteers. The process of manually tagging (segmentation) and annotating place names in text with entries (toponyms) from a geographic gazetteer, here called ‘Geo-Annotation’, is laborious, costly, time-consuming and error-prone. The scarcity of publicly available geo-annotated corpora can partially be attributed to the lack of available efficient software infrastructure capable of facilitating this laborious task.” [46]
-“Extracting the ‘correct’ location information from text data, i.e., determining the place of event, has long been a goal for automated text processing.” [47]

2.2.9 Event extraction

-“Event extraction is the process of recognizing defined event types in text (e.g. “attack” or “protest”) and extracting and classifying the actors involved in the events.” [48]
-“To be useful in subnational research, these events require information on the location where they occurred. A second related information extraction task is ‘geoparsing’, the process of recognizing place names in text (‘toponym recognition’) and resolving them to their coordinates or gazetteer entry (‘toponym resolution’). [48]
-Tokenize a sentence. Potential for multiple events per sentence, potential for zero, one, or multiple tokens per event. Tokens labeled as 1 (if location of event) or 0 (otherwise). Assumes “that events are ‘anchored’ by a verb, is a common assumption in semantic role labelling, a closely related task to event-location linking”. [48]
-“Many existing open source geolocated event datasets, including GDELT and Phoenix, make no effort to explicitly link events and locations, simply returning a top location from a sentence, without using information on the extracted event to inform the geolocation step, which has also been used in NLP”. [48]
-“Primary focus location” of Imani (2017): “makes the simplifying assumption that documents have one single, fixed ‘focus location’ that is invariant to different potential events in the document.” [48]
-“While humans are able to pick up on nuance and deal with grammatical complexity that machines still cannot handle, humans are also unsuited to the tedium of labeling thousands of sentences and may be susceptible to drift in their definitions or understanding of the task. Not only is the automated method vastly cheaper and faster than the human process, it does so with accuracy at least as good.” [48]
- “Text also holds a great deal of factual information and new techniques are needed to allow researchers to extract political information from text.” [48]
-“Many quantitative studies of conflict rely on event data. Recently, these studies have also retreated from the country-year framework and have focused on disaggregating the event flows both in terms of space and time. DIsaggregating temporality - even to the

daily level - is a straightforward task. But figuring out precisely where an event actually occurred is a difficult and uncertain task that has been perplexing for most contemporary event data efforts". [47]

-"Most data in the conflict realm comes from non-official sources. FOr many that means some form of data collected form historical and journalistic sources. This need is often filled by event data, which are typically collected on a daily basis, and can be aggregated temporally to the level required by the analysis. Even data can also be aggregated to the geographical region that is appropriate. Given the increasing demands fro event data, the scientific community has recently devoted significant efforts to automate the data collection process. Having humans read and code a large set of archive documents sometimes limits reproducibility, and hence hinders scientific research. It is also expensive and limits the currency of the data. Further, ensuring inter-coder reliability is challenging, especially over global events that span decades." [47]

-"These automated event data allow researchers to observe and extract information on politically relevant events around the world in near real-time." [47]

-"Outstanding issues [in automated data collection] include machine translation of texts in foreign languages (wherein great progress is being made in both Chinese and Arabic), duplicate reports from multiple sources, and the relatively low accuracy in determining the event location." [47]

-"Named entity recognition in the context of geolocation involves determining which words in the given sentences are location names. In principle, the task of capturing location names from texts can be done easily by using a dictionary. In practice, however, developing a dictionary that is sufficiently comprehensive for such a task may be challenging. To begin, the geographic boundary of the texts being analyzed may be unclear, given that the domain of many even data is the entire world. Further, because conflict events often spread to new and rural places, texts may include location names not defined in the gazetteer. Still further, a location name may be written in multiple forms, requiring the dictionary to comprise every variant for each location." [47]

-Ground truth: hand coded set of actual locations for training and verification [47]

-"The task of determining event locations involves three steps, each non trivial." 1. named entity recognition (NER): "all location names are identified and extracted from an appropriately preprocessed text. This step is a prerequisite for the other steps because to determine the location of an event in a news article, capturing the exhaustive list of location names is required." 2. Ambiguation/resolution: "which involves identifying the actual location of the recognized name string. Once this is accomplished, it is possible to extract the ontologically defined meaning from the text in terms of who does what to whom, and when and where". 3. Determine if disambiguated location names are the event location. [47]

-"geolocating events (identifying the location of the event described in a document) is an objective for many scholars, particularly those who intend to collect and build original databases from text corpora, be they news articles, congressional records, campaign speeches, party constitutions, or twitter feeds. While automating this task will aid many, the research avenue in this topic is still under development." [47]

-"Given that a substantial number of location words are incorrect event locations, the automated event data community needs a better coding scheme that can reduce the error rates." [47]

-Event-relevant: "those locations that are part of the main description of the event of interest, i.e., all locations that are key to the narrative of the event of interest." [47]

-Event-occurring: "all locations where events occurred regardless of whether the event is

the event of interest.” [47]
-Eventirrelevant and event occurring location: “such could occur when the raw texts contain news summaries of events that are not of interest.” [47]
-“An N-gram is a sequence of N words. Collections of N-grams are known to provide valuable information about each word in a phrase, taking into account the complexity and long distance dependencies of languages... Given that the collocation patterns in which the event-occurring location words appear differ from those of the non-event-occurring collocations, teh N-gram patterns are able to provide the contextual information of even-occurrence to our classifiers.” [47]
-“Geolocation inference at the event level estimates the location of events mentioned in text. This level of inference predominantly relies on geoparsing the process of identifying geolocations in text and disambiguating between multiple toponym references... event geolocation inference might not reflect the actual location of individual tweets.” [37]
-User geolocation: “user locations can be predicted by utilizing toponym references within their tweets a well as user metadata such as friend networks and time zones.” [37]
-Tweet level (article level): “estimates the location of individual tweets. THis differs from user-level prediction in that a tweet might be posted in a separate location from where they live, such as during a vacation or work hours.” [37]

Examples:

- Cliff-Clavin
- Mordecai
- Stanford-CoreNLP
- BERT: BERT-based language model. Google’s Bidirectional ENcoder Representations from Transformers, “powerful, pre-trained deep learning based language models” [41]. -BERT: “a general purpose language representation model pre-trained on millions of articles on English Wikipedia an BookCorpus.” [41]
- Wheb-a-where: Web-a-Where: “This small size imposes a serious limitation on Web-a-Where’s practical geotagging capabilities, as it is unable to recognize the small, highly local places that are commonplace in articles from local newspapers.” [44]

Tumba

- Tumba: a web search engine for Portugal (Gomes & Silva 2003) [1]
- “We propose to further improve the quality of search systems, by integrating the geographical knowledge that can be inferred from web resources.” [1]
- Scope: “the region, if it exists, whose readers find the page more relevant than average. A geographic scope is specified as a relationship between an entity on the web domain (a HTML page or a website) and an entity in the geographic domain (such as a location or administrative region). The geographic scope of a web entity has the same footprint as the associated geographic entity.” Geographic scope definition for reader engagement (not generically organizing location for research or general applications later [1]
- “Geographic information is pervasive on the web. An analysis of 3,775,611 pages 8,147,120 references to the 308 Portuguese municipalities (administrative division of

the territory corresponding to populated sub-regions of the Eurstat NUT 3 area), an average of 2.2 references per document." [1]

- Geo Tumba provides a retroactive organization and characterization of existing webpages, versus specifically defined ruing writing. [1]
- "The Portuguese ontology included more place names, but results in terms of recall are inferior to experiments using hte smaller global ontology... results indicate that recall does not improve considerably with the amount of available place names." [1]
- "Our framework differs on the emphasis put on geographic name entity recognition, the use of a graph ranking method for assigning a single scope to each document, the extensive use of names instead of coordinate information, and the availability of ontologies associating entities to geographic scopes." [1]

LocateXT
-Search unstructured data for spatial locations [49]
-Geneartes point features representing locations [49]
-Attributes include file, context related text, dates, keywords [49]
-Not addresses [49]
-Custom locations: "When documents are scanned, they are examined for place names specified in a custom locations file. The custom locations file associates a place name with a spatial coordinate. A point is created in the output feature class to represent each location found." [49]
-fuzzy matching: applies to custom locations. 70% of alphanumerics match [49]
-Require word breaks setting "when word breaks are required, text is considered a word when it is bounded by whitespace or punctuation characters as in European languages." [49]
-Require word breaks setting "when word breaks are required, text is considered a word when it is bounded by whitespace or punctuation characters as in European languages." [49]
-Custom location: "is defined by associating text with a spatial coordinate. Current and historical places, and natural features and structures can all be defined as custom locations in a custom location file (.lxtgaz)." [49]Input your own gazetteer (custom or otherwise). Again, only points (not polygons).

-Integrated Crisis Early Warning System (ICEWS) [47]
-Open Source Event Data Alliance [47]

GeoTxt:

- a scalable geoparsing systems [43]
- SPIRIT: spatially aware information retrieval on the internet [43]
- Geographic footprint of documents [43]
- Heuristics for toponym disambiguation: spatial proximity ("spatial minimality"), co-occurence (relation based on spatial hierarchy levels) [43]
- Appropriate for hyperlocal applications: "advantageous on a corpus with a smaller known geographic footprint" [43]

DeepGeo [37]

2.2.10 Gazetteers

-“The locale name for an object within a language area is termed an endonym. Outside this language area, te ame object may have other names according to the respective language. These variations are called exonyms.” [38]
-“The United Nations Group of Experts on Geographical Names (UNGEGN) is run by specialists from the fields of linguistics, cartography and history. The UNGEGN requests national gazetteers (alphabetical lists of names, with coordinates and other data) to promote the use of nationally standardised names on maps and in written documents.” [38]
-”An exonym is a place name that isn’t used by the people who live in that place but that is used by ohers.” [50]
-”a locally used toponym - that is, a name used by a group of people to refer to themselves or their region (as opposed to a name given to them by others) - is called an endoynm. [50]
-Gazetteers assist in geographical entity identification performance [1]
-“The larger the gazetteer (in terms of coverage and detail), the harder the task of toponym resolution, since the gazetteer will contain a higher number of ambiguous names. However, a large gazetteer such as GeoNames anables resolving more places to toponyms.” [43]
-Progressive geo-coding environment: “this configuration of geo-referencing system ensures that a local gazetteer is incrementally enriched through local community use and the demand for the effort of human go-coding is kept to the minimum.” [32]
-“Gazetteers are well-structured and maintained, but they have limited coverage on those less known places at local levels. GIS databases have geographical features in greater geometric precision and details, but they tend to have limited textual metadata for matching to place names. Open Source geographical information now comprises and ever growing part of geographical knowledge, and they frequently include vernacular descriptions of locations, as well as references to imprecise areas. The problem with using CGI is that their local coverage is uneven from locality to locality, and they tend to have various degree of quality and trust. Ideal VGI is supposed to be produced entirely by individual citizens that reside in their locality, and therefore potentially represent the best knowledge about their local.” [32]
-“Majority of gazetteer entries use points (coordinates) to represent areas (such as a city), which is considered inadequate for local features. [32]

Gazetteer examples:
-OpenStreetMap: open data, licenced nder the Open Data Commons Open License (ODbL) by the OpenStreetMap Foundation (OSMF): [51]. Relies on OSM “performed the best”... “DBpedia is not focused on geographical information; therefore, it does not contain the metadata useful for the system’s future use (e.g., extents and full addresses). Also, OSM has more fine-grained locations and more accurate geo-coordinates than Geonames”. [29]. Use [29]
- Geonet Names Server (GNS): US NGA maintained: GeoNames. “The GEOnet Names Server (GNS) is the official repository of standard spellings of all foreign geographic names, sanctioned by the United States Board on Geographic Names (US BGN). The database also contains variant spellings (cross-references), which are useful for finding purposes, as well as non-Roman script spellings of many of these names. All the geographic features in the database contain information about location, administrative division, and qual-

ity. The database can be used for a variety of purposes, including establishing official spellings of foreign place names, cartography, GIS, GEOINT, and finding places." [52]. Use [29]; "The GeoNames gazetteer was chosen for GeoText because of its extensive coverage, quality, inclusion of metadata (such as alternate names and geographic hierarchical information),and frequent updates". [43]; Official sources include U.S. National Geospatial Intelligence Agency (NGA) and the U.S. Board on Geographic Names, "assuring extensive coverage and quality throughout the world." [43] "GeoAnnotator uses GeoNames as its default gazetteer because of its frequent updates, ease of ability by years to correct information, extensive coverage and quality and inclusion of metadata items necessary for geoparsing/geo-annotation, such as alternative names and spatial hierarchies. All GeoNames toponyms have a unique identifier, enabling creation of corpora useful for linked data applications." GeoNames: "the richest gazetteer available at the time of this writing." [46]
-DBPedia [29]
Requires premium subscription for polygon access. -Getty THesaurus of Geographic Names (TGN): $http : //www.getty.edu/research/conducting_research/vocabularies/tgn/$ "TGN is a structured vocabulary including names and associated information about both current and historical places around the globe." [1]

2.2.11 Local lexicons

-Automatic inference of local lexicons: "1) Stability: a Local lexicon is constant across articles from its new source. 2) Proximity: Toponyms in a local lexicon are geographically proximate. 3) Modesty: A local lexicon contains a considerable but now excessive number of toponyms." [44]
-"a spatial lexicon can be classified as a local lexicon if and only if the toponyms within it are geographically proximate." [44]
-"Associating a single local lexicon with each data source allows for a variety of applications. However, it may be possible to fine-tune the use of spatial lexicons in situations involving different types of content. For example, a blog may track several different topics simultaneously, and use different spatial lexicons for each oipc. Furthermore, individual authors may write for specific audiences as well, as in the case of journalists stationed in certain geographic areas and concentrating on stories in that area. It thus might be beneficial to determine separate spatial lexicons assumed by different authors, and further improve geotagging performance. Ore generally, we might associate a particular spatial lexicon with any type of entity found in each document, be they authors, persons, organizations, or particular keywords. For example, upon finding a mention of 'Robert Mugabe', we might assume a spatial lexicon including SZimbabwe and nearby locations, even without specific mentions in the text." [44]Complicated yet potentially effective alternative to manual association. [44]
-"It would also be interesting to detect and observe evolving spatial lexicons over time for data sources with evolving geographic interests, thus further improving geotagging on these sources. For example, the first few articles of an ongoing, prominent news story will often fully specify the toponyms relevant to the story. Later articles in the series, however, will often underspecify the same toponyms, since they have already been introduced into the audience's spatial lexicon and been fully resolved in early articles." [44]
-"As newspapers and other data sources continue to move into the virtual space of the INternet, knowing and using spatial lexicons will be ever more important. Previously localized newspapers will cater to a broader, global audience, and thus will adjust their notion of their audiences' spatial lexicons, perhaps limiting or ding away with an assumed

local lexicon altogether. On the other hand, as more and more people publish highly individual and geographically local content, inferring individual local lexicons will be a necessity for correct geotagging. Geotagging with knowledge of local lexicons will thus continue to play a large role in enabling interesting geospatial applications. [44]

2.3 News localization

- News is a distribution of information to humans which in turn affects their actions. We have manners of measuring the physical world via sensors, and now are layering in emotional responses. The news is a feedback loop that is at once an input to human decision making and a report of some of those outputs.
- "Everything happens somewhere and some-whence. Spatial and temporal aspects of data lead to critical insights into the information contained in it." [26]
- "in the current hypermodern context, a smart city must also be able to identify the main components of an event (where, when, who, what, how), to analyze it, to provide location-aware (and contextual) sense to it and, to react (actuate) properly (and in real-time - at least compliant with the nature of this event). [3]
-"Spatio-temporal aspects of data lead to critical information." [26]
-"A media or news desert is an uncovered geographical area that has few or no news outlets and receives little coverage. Mapping locations mentioned in news articles is the primary step in identifying news deserts." [41]

2.3.1 News evolution

-"There's a lot of opportunity for innovation across the news ecosystem, and this is also true for local publishers... They're connecting the dots with their communities and are trying to build innovative and more engaging experiences using local data." LB [19]
-"Having the tools to access and analyse this data is now critical for effective reporting... There are many powerful public interest stories out there that will only be discovered if traditional investigative techniques are combined with technology" Megan Lucero [19]
-Linked in provided a business skillset "which you don't often get in journalism because we're, you know, so adamant about separating the editorial and commercial side of things." Now this is more present in journalism (as of the past 5 years). [53]
-"I really believe in the community element of journalism. So I'm really interested in not just broadcasting but hearing back and bringing people together. It not where I want it to be, but I think by calling it a community I'm stating that intention." [53]
-Matching product to audience: "You have to throw a lot of lines out in the water and see which ones catch. So you end up having a lot of platforms." [53]

2.3.2 Personalization

-"While personalization of content is recognized as a powerful advantage for publishers and readers alike, it can mean that people aren't often exposed to a range of viewpoints on a particular subject." [19]

2.3.3 Blogs

-“Blogs, as social software, are designed for rapid content creation, archiving, and syndication within online communities. Due to the community intelligence, blogs have lately become a valuable source of information. Blogs, used as collaboration environments, support people in associating tags with content that they generate, share, or consume within a community. More generally, by tagging content in the web, one makes this content shareable to other web participants, as well as links it to other web contents.” [27]
-“There has been very little research on geoparsing microblog text.” [43]

2.3.4 Indie journalism

-Indie journalism: ”independent side-projects” as entrepreneurial endeavors of journalists. [53]

2.3.5 Local journalism

-“The news industry may be in flux, but it’s vitally important that good local journalism doesn’t disappear.” Megan Lucero, Director at The Bureau Local [19]
-“Local journalists play an important role in democratic society by holding power to account. However, newsrooms have fewer resources to dedicate to much-needed investigative reporting.” [19]
-“While almost every reader survey concludes that local news is the most highly valued of all content, making it pay is another matter entirely.” [19]
-“The appetite for hyperlocal news is vast” [19]
-“The core problem in local and hyperlocal news is that there is no sustainable business model,” says Gabriel Kahn. “These areas have no voice, they have no way of discussing problems in the community and, whether it’s governments or companies, they have no accountability.” [54]
-“Local news is a medium for communication among residents, and they reflect the nature of local geographical knowledge. THe residents’ local knowledge concerning their living environment is often invisible, descriptive and vague, and thus difficult to collect.” [32]
-“Local newspapers can be considered as one important tool to support the accumulation and sharing of local knowledge.” [32]
-“Because local news serves the important role of being repositories of political, social, and cultural knowledge in the community of practice, it is important to make such knowledge accessible to community members who can maintain the awareness of what the community have known or done in the past when facing a new or repeating problem.” [32]
-“citizens’ attention to local news promoted political participation.” [32]
-“Local news are textual artifacts of human experience on the places they live and interact with, and they play an important role in the making of of place meaning. They reflect what local people think and are written for the local people” [32]
-“Hyperlocality is a concept that has emerged in both online journalism and in mobile advertising circles to designate community-oriented digital content endemic to small geographies such as neighbourhoods, streets, or single postcodes.” [31]

2.3.6 Georeferenced News

-“Once a blog system organizes the entries with geographical semantics and display localities, spatial analysis can be performed based on the user-tagged information. In order to develop such system, the domain specific ontologies that take geographical knowledge into account should be built.” [27]
-“In the era of web 2.0, blog systems such as GeoBlog provide an approach for individuals to input variou geographical data and the asynchronous functions to meet the asynchronous interaction requirements. This apparently widens the data source of geographical information applications. [27]
-“Considering the map as an object that conveys meaning, we had to recognise not only the agency of the participants but the agency of the object itself. Yes, this mapping of experiences has the potential to be a powerful source of information, and it also has the power to label places as safe, dangerous, racist, oppressive, accepting, and so forth.” [12]
-“Participants and researchers come with assumptions wrapped in powerful collective memories. Acknowledging them is important, but we also need to look for disturbances and consider the impact of those disturbances.” [12]
-“If you are a journalist for a news outlet where location is a key part of the story (such as a local or hyperlocal news site), it is worth geotagging your stories to help search engines deliver local results:” [55]
-“Location-based feeds have huge potential. They may further develop as a way Google and other search engines return results based on a person’s location; they may be used by social newsreader apps as a way of delivering [55]
-“Political news reports are populated all over the world in various languages It has a great value to automatically detect the geolocation from these reports for better understanding of the associated events.” [30]
-“Primary focus location as the actual location where the event occurred amongst other focus locations mentioned in the report.” [30]
-“determining crime pattern locations, predicting the place of protests and political unrest, and identifying the geolocation of natural disasters. Such applications can largely benefit form identifying precise geolocation information in a timely manner to provide better support for decision making.” [30]
-“News articles contain a wealth of implicit geographic content that if exposed to readers improves understanding of today’s news.” [42]
-“Given that much of the interest in news is motivated by location-related attributes of readers (e.g. where readers are situated, ahil from, aspire to be), it is somewhat surprising that they cannot deal easily with the two most common types of spatially-related queries: 1. Feature-based - ‘Where did story X happen?’ 2. Location-based - ‘What is happening in Location Y?’” [42]
-The traditional newspaper layout “is linear and static, whereas the ap interface is dynamic, in that the articles associated with a particular location can vary over time without disturbing the positioning of other articles.” [42]
-“an audience’s local lexicon plays a key role in how news authors write for their audiences.” [44]
-“geo-referencing local news can contribute to improved accessibility of community knowledge.” [32]
-“For newspapers targeted at local communities, geographical locations are likely to be in finer granularity and demand gazetteer-based approach” (vs. statistical language model approach) [32]
-“Due to uniqueness of each locality, there is a lack of local gazetteers that reflect the

richness and details of local spatial language." [32]

-"Location ambiguity in the local has its own unique nature that presents both challenges and opportunities to resolve them." [32]

-In this study, only 16 percent of references could be mapped without ambiguity, 59% were discernible with the applied heuristics, and the remaining 12% "rely on human annotation to create footprints." [32]

-Therefore, the gazetteer used for geo-referencing local newspaper articles should be place-specific." [32]

-"What kinds of spatial language are used in the locality? Answers to this question will make clear the nature of gazetteers needed for geoparsing local newspaper." [32]

-"Local news articles are an important source of knowledge about local events, place-specific culture, and peoples' thoughts about their environment. Reliable geocoding of such articles is the first step towards unlocking such local knowledge for community engagement and development." [32]

-"existing geo-referencing methods and tools do not work well for local news because they do not reflect the ways local people encode and communicate geographical knowledge." [32]

-"FOr human coders, locating events by reading a news article may be time-intensive, but straightforward. This is not the case for machine-coding: many news articles contain multiple location names, such as the location of the journalist writing the story, the birthplace of a person being interviewed, or the place of a similar event that occurred several decades ago; at times, human names are identical to geographic names,; and location names are transliterated into English in a variety of potentially confusing ways. All these sources of noise in the data increase the difficulty in automatically locating events." [47]

-"Further complicating matters, new articles often use nearby landmarks to indicate the location, in lieu of using the official names." [47]

2.3.7 Automated location extraction from published corpora

-"The networks are global and spatial proximity is no longer a determinant of relations established in communities, as it used to be. These new communication models allow us to be in the world, however we run the risk of not reaching our street, or our city, and in it to see the reflection of the image of what we desire our urban environment to be." [24]

-"To map local news coverage, it is important to extract precise location mentions from textual news content." [41]

-The traditional newspaper layout "forces readers to perform a brute force sequential search (i.e. read the various articles while looking for mentions of the locations which interest them)." [42]

-"Determining the geographic focus of a document can be challenging, as not all documents have an easily identifiable focus, and not all locations referenced in a document may be related to its focus. For example, news articles often contain the address of the newspaper that published the articles." [42]

-"One related issue is that stories may continually change and be updated, even after they have been 'published' in an RSS feed." [42]

-"As retrieved, the story webpage is unsuitable for article processing, as it was meant to be read by humans and contains extraneous formatting markup and rendering scripts.... Furthermore, the extraction must be independent of the source website, as it is infeasible to create custom extraction rules for each individual website." [42]

-"The successful execution of location-based and feature-based queries on spatial databases requires the construction of spatial indexes on the spatial attributes. This is not simple

when the data is unstructured as is the case when the data is a collection of documents such as news articles, which is the domain of discourse, where the spatial attribute consists of text that can be (but is not required to be) interpreted as the names of locations. [44]
-"news articles (and more generally, documents on the Internet) are written to be understood by a human audience, and therefore geotagging will benefit from processing (i.e., reading) the document in the same way as an intended reader)." [44]
-"the reader's spatial lexicon - those locations that the reader can identify and place on the map without any evidence - is very limited. In fact, even more importantly, this inherent limitation means that a common spatial lexicon shared by all humans cannot exist, which is one of the key principles used by systems such as MetaCarta and Web-a-Where." [44]
-"the existence of a reader's local spatial lexicon or simply local lexicon that differs from place to place, and that it is separate from a global lexicon of prominent places known by everyone." [44]
-"The local lexicon is even more necessary when geographically indexing locations with smaller spatial extent which correspond to address intersections. . . , since street names are even more ambiguous than regular toponyms." [44]
-"While statistical NER methods can be useful for analysis of static corpora, they are not well-suited to the dynamic and ever-changing nature of the news" [44]
-"The main idea behind our geotagging framework within an article, to make it easier for human readers in the author's intended audience to recognize and resolve toponyms. Authors create this framework by using linguistic contextual clues that we can detect using heuristic rules. Furthermore, readers are expected to read articles linearly, so article language has a contextual and geographic flow. Toponyms mentioned in a sentence will establish a geographic framework for subsequent text." [44]
-"Certain phrases in article text denote relative geography, which is language that defines a usually imprecise geographic region in terms of distance from or proximity to another geographic location. These imprecise regions are important because they usually target the geographic areas where the events in an article took place, and therefore are useful for resolving the article's toponyms. . . We refer to the tononyms in such phrases as anchor toponyms, and we term the resulting regions as target regions." [44]
-Recall drops "reflects the fact that most gazetteers are still rather incomplete or at least not in sync with the frequency of use of location descriptions that do not have formally defined boundaries, such as 'New England' and 'Upper West Side'." [44]
-"Event extraction from news articles is a commonly required prerequisite for various tasks, such as article summarization, article clustering, and news aggregation. Due to the lack of universally applicable and publicly available methods tailored to news datasets, many researchers redundantly implement event extraction methods for their own projects." [45]
-"The extraction of a news article's main event is an automated analysis task at the core of a range of use cases, including news aggregation, clustering of articles reporting on the same event, and news summarization." [45]
-"Beyond computer science, other disciplines also analyze how news outlets report on events in what is known as frame analyses." [45]
-"Explicit event descriptors are properties that occur in a text to describe an event, e.g., the phrases in an article that enable a reader to understand what the article is reporting on." [45]
-"State-of-the-art methods for extracting events from articles suffer from three main shortcomings. First, most approaches only detect events implicitly, eg.g. By employing topic modeling. Second, they are specialized for the extraction of task-specific properties, e.g., extracting only the number of injured people in an attack. Lastly, some methods extract

explicit descriptors, but are not publicly available, or are described in insufficient detail to allow researchers to reimplement the approaches.” [45]
-“Similar to temporal phrases, locality phrases are often heterogeneous, i.e., they do not only contain temporal NEs but also function words.” Nomanatim: geocode leveraging OSM [45]
-“For ‘when’ and ‘where’ questions, we found that in some cases an article does not explicitly mention the main event’s date or location. The date of an event may be implicitly defined by the reported event, e.g., ‘in the final of the Canberra Classic.’. The location may be implicitly defined by the main actor, e.g., ‘Apple Postpones Release of [...]’, which likely happened at the Apple Headquarters in Cupertino. Similarly, the proper noun ‘Stanford University’ also defines a location.” [45]
-“The journalistic 5W1H questions are capable of describing the main event of an article, i.e., by answering who did what, when, where, why, and how.” [45]
-“Answering the 5W1H questions is at the core of understanding any article, and thus an essential task in many research efforts that analyze articles.” [45]
-Class imbalance: “often degrades machine learning approaches, which skews the classification capacity in favor of the most crowded classes.” (applicatin to tags) [10]
-Most previous studies are english based [10] innapropriate for local investigations in other langauges

2.3.8 Examples

GLOBAL

GDELT Review Some projects are already mining place (as well as other attributes) from existing data lakes of publication data to provide geospatial and temporal distributions. One such effort is The GDELT Project, which extracts place as well as actors, sentiment, and event connection (among other elements) from journalistic media across the globe, including publications from as far back as 1979. This and similar projects are powerful and hugely informative, especially as they apply to existing published data. The proposed project should leverage such tools for the inclusion of historic data into the developed database for investigation into the past (already published) incidents. However, the existing automated extraction includes several challenges:

1. It is not yet perfect, and places may be misattributed (Lisbon, Ohio in the USA may be accidentally attributed to Lisbon, Portugal).

2. It does not support the subtlety of incidents occurring in non-conforming places (an incident may not apply to a single administrative boundary but really fall into a subsection of one or several).

3. It requires technical prowess and tools to explore the data. A user is unable to define a spatial area of interest (such as their route to work with a half mile buffer or some other irregular shape) and search for all spatially related results, nor it is easy to apply temporal or thematic attributes without prior experience querying results.

Therefore, this project offers a functionality specific to the defined user types of news publication services and provides an appropriate user experience to these.

5W1H -“Since events according to our definition occur at a single point in time, we only retrieve datetimes indicating an extract time, e.g., “yesterday at 6pm”, or a duration, e.g., "yesterday", which spans the whole day." [45]

Rivera2020
-Tasks: "1. Get a corpus of news published by the local newspaper through an RSS (Really Simple Syndication) feed, 2. Get a vector characterization based on the well-known Bag-of-Words (Bow) representation, 3. Select features through a mutual information-based method, 4. Train a supervised-learning model, 5. classify online news reports in real-time; the interest is in the 'traffic accident' class, 6. process the text of the RSS reports to retrieve the location where the accidents happened, and 7. notify users about the events on a map." [10]

Lee2019 -Automatic POS extraction: Stanford named entity recognizer, apache open NLP algorithm, MIT information extraction (MITIE), OEDA [47]
-Leverated location lists from GEonames, Wikipedia, and Google map
-Preprocessing: building gazetteers and import of data
-Follow for verification and comparison (Validation) type steps if there is time after

LOCAL

Crosstown
-"Our MIssion: Crosstown seeks to deliver community-level data and analyses to the people of Los Angeles who want to make their neighborhoods and city safer, healthier and more connected." [56]
-"We want to turn those numbers into stories that people can use to understand what's happening around them:" [56]
-Organized by neighborhood boundaries (LA Times), different from administrative boundaries. [56]
-Leverages public health, crime, administrative inputs, etc. [56]
-Not geolocating news, but building stories based on location [56]
-"Crosstown's focus is not on increasing revenue but decreasing costs of producing local news. It does this through "mass customization" of public datasets in a three-tier system". [54]
-Geocoded and aggregated by neighborhood [54]
-"We have created hyperlocal news at scale, and also allowed people in the city to see how their neighbourhood fits in with others and understand their experience in context" (- Gabriel Kahn) [54]
-"Tier one: data collection informing news stories". Collecting and geo-locating "quality of life" data. "Journalists can then access the data and spot newsworthy trends at a glance." "Tier two: data visualisation and dashboards" "There are a lot of different opportunities to take that data and not simply turn it into a classic news story, but turn it into other opportunities for the audience to engage with it." "Tier three: neighbourhood newsletters". Avoid the need to have a dedicated reporter in each area. [54]

- NewsStand

- "NewsStand extracts the 'interesting' phrases that are most likely to be references to geographic locations and other entities by using NER methods. LOCATION phrases are stored as geographic features of the entity feature vector. Then, it uses a Gazetteer to find those geographic features in the entity feature vector that are names of actual locations. It also employs Gazetteer to identify the hierarchical information for each location (i.e. country and administrative subdivisions).. After that, it extracts geographic focus (or fous location) based on the frequency of the locations in the news." Description of NewsStand used specifically for identifying primary locations of news articles? [30]

- "NewsStand monitors RSS feeds form thousands of online news sources and retrieves articles within minutes of publication. It then extracts geographic content from articles using a custom-built geotagger, and groups articles into story clusters using a fast online clustering algorithm. By panning and zooming in NewStand's map interface, users can retrieve stories based on both topical significance and geographic region, and see substantially different stories depending on position and zoom level." [42]
- NewsStand: Spatio-Textual Aggregation of News and Display [42]
- Based on "transactional database technology" [42]
- "It places markers representing story clusters on an interactive map interface, thereby allowing meaningful, visual exploration of the news." [42]
- "The interplay between significance and zoom level is an important feature of NewsStand, and differentiates is greatly from existing spatially-referenced news reading systems (e.g. the Reuters News Map, the maps locations found in stories using MetaCarta)." [42]
- Global scalability: obscures articles not of interest to a global audience. [42]
- Geographic information extraction: "1.Provider scope, the publisher's geographic location; 2. Content scope, the story content's geography; and 3. Serving scope, based on the readers' location." [42]
- A global solution [42]
- "In NewsStand we are also interested in the geographic focus of a collection of news articles about the same subject/topic, rather than just one article, and this is done with the aid of a document clustering algorithm." [42]
- "After a new article has been introduced to the system, NewsStand must locate and extract the geographic content from the article. This process, described earlier as geotagging, unifies the explicit textual article content with the implicit geography, and enables spatial exploration of the news." [42]
- "Our main goal in designing NewsStand's user interface was to convey as much geographic and non-geographic information abou current news as possible. The interface consists of a large map on which stories are placed, and the viewing window serves as a spatial region query on the geotagged news stories. Users interact with NewsStand using pand and zoom capabilities to retrieve additional news stories. As users pan and zoom on the map, the map is constantly updated to retrieve news stories for the viewing window, thus keeping the window filled with stories, regardless of position or zoom level." [42]
- "NewsStand also features a smaller map that shows the geographic span of the selected story. This minimap allows users to easily see the selected story's geographic

focus, without having to leave the area of interest on the main map." [42]

- To avoid geographically clustered visualization of uneven news coverage "Marker selection is therefore a tradeoff between story significance and spread. "To achieve a balance in marker mode, NewsStand divides the viewing window into a regular grid, and requires that each grid square contains no more than a maximum number of markers. The markers to display are selected in decreasing order of story significance and story age. This approach ensures a good spread of top stories across the entire map. [42]

Suntimes Crime
-mapped instances of crime within a city [57]
-"Find out about crime in your neighborhood and your city. The best and most timely analysis of crime trends in the city of Chicago." [57]
-No map or spatial exploration of the news. No textual/tag organization of place.

In Your Area
-InYourArea lets you follow the latest local news, information, events and more in your area. InYourArea covers local news and information for all Towns, Cities, Villages and Hamlets in the United Kingdom. As well as the latest news and information for where you live, you can also connect with other members of your community, submit events, promote a local business and more." [58]
-"InYourArea includes news from all the top local and national news publishers and blogs. It includes updates from local councils, police and public services." [58]
-"InYourArea covers news for all UK Cities, Towns, Villages and Hamlets including London, Birmingham, Manchester, Leeds, Liverpool, Sheffield, Cambridge, Bristol, Norwich, Reading, Cardiff, Edinburgh and more. So download it today and see what's happening in your area. [58]
-Post code based (not searchable by specific locations) [58]
-Method for sharing relevant information by postcode [58]
-Signup: input post code, select three areas of interest [58]
-Includes: state of transit (roads, rail, bus, TFL); weather; notices; recent updates (comments and news stories); all news vs. my news; ability to interact (like, comment, share, follow); no map searching options; whats happening (discussions, about my area, planning applications, funeral notices, council notices) and find (homes near you, items for sale, deals and offers, things to do, local services) [58]

2.3.9 Manual location extraction from published corpora

-Finding 1"There are significant portion of place names in local news that are vernacular, vague places, or finer granularity places that were not found in gazetteers." [32]
-FInding 2: "THe majority of those place names found matches from gazetteers are found to be ambiguous." [32]11 -Finding 3 "For those place names that were not able to find proper footprints in gazetteers, humans can create footprints with ease when supported well." "Because GeoAnnotator helped human in locating the area and providing good spatial context, drawing the footprint in the map view is relatively straightforward." [32]
-Finding 4: "There are high degree of verbatim repetition of place references across local news articles over time." potential "to end up with a local gazetteer that is enriched over

time". [32]
-"heuristic rules that human coders use in disambiguation of place references in local newspaper are very different from those used in resolving ambiguities in global news." [32]
-"generic gazetteers are not adequate for geoparsing and geomatching for local news." [32]
-Suggestive geocoding: "Geocoding... requires deep, human-level knowledge to evaluate a potentially large list of candidate matches and make a choice. THese tasks re best done by bringing human and computer to work collaboratively. A simple way to do this is to have a computer doing step 2 (generating all the candidate matches) and uman take care of step 3 (disambiguation). Alternatively, computer can further reduce human efforts in step 3 by ranking the candidates such that those with higher degree of likelihood are placed on the top of the list, and those candidates that are deemed impossible after considering the context are eliminated from the list." [32]
-"While a system that follows the progressive geospatial referencing framework is expected to enrich its gazetteer and improve its performance over time, ti does require human effort to deal with uncertainties, and complex geographical descriptions that have no gazetteer match (e.g. a long prepositional phrase representing a place). THe key challenge here is to reduce the human effort to the minimal and accelerate the process of incremental improvement." [32]
-"Geo-annotation requires either local knowledge of the place or general geographic knowledge and the initiative and ability to further research the existence and location of places." [46]
-"Presenting place names to users with pre-annotation hastens the geo-annotation process but comes with a risk. It is likely, especially for larger pieces of text, that users might miss place references that are not already highlighted in text by the pre-annotation process, and therefore, these entities are not added to the map view and might be omitted due to visual absence." [46]
-"the quality of a geoparser's output annotation is domain-dependent. Depending on the kind of text (e.g., news story, social media posts or academic articles), the geographic coverage area of documents (e.g., North America, Middle East, Eastern Europe), and the geographic level of toponyms (e.g., country level, state level, physical features, landmarks or building names or addresses), the geoparser may underperform and therefore create additional work for users performing annotation." [46]

EXAMPLES:
GeoAnnotator Workbench
-Protocol: "1. Open an article and read through the whole article to understand the story; 2. Geoparsing: identify all place names (or prepositional phrases) and highlight them in the article; 3. Geomatching and disambiguation: for each identified location reference, search the best gazetteers to find all matches, compare candidates, and finally make a choice from the candidates, and finally make a choice from the candidates. In case no candidate match is found or all candidates are rejected by the human analyst, he or she can create a new footprint and add it to the local gazetteer. 4. Coding: depending on the gazetteer matching outcomes and the disambiguation strategies used by the analyst, each location reference is coded by a case number." GeoAnnotator workbench [32]
-Uses a local gazetteer (subsect of nominatim), nominatim global, and google maps [32]
-"1. Start with a VGI-based gazetteer that is best for the target locality. The quality of this gazetteer does not matter so much, as it only provides a starting point for us to bootstrap the system. In the current implementation of GeoANnotator+, twe bootstrap the system by taking Nominatim as the initial source of gazetteer. Nominatim provides

a search engine API for using OpenStreetMap data as a proxy gazetteer to match place names. The quality of the footprints in Nominatim is quite good, as it normally returns matches of local features by good approximation of their real shape (points, lines, or polygons). We choose Nomination as the initial gazetteer because it is considered the best VGI-based geographical data sources with good coverage in most localities worldwide. 2. Progressive geocoding to enrich gazetteer incrementally. Provide a semi-automated workbench for human analysts to evaluate all matches and make a choice. If none of the matches are appropriate, analysts will create the footprint in the workbench. The system not only remembers the result of this human-generated geocode, but also adds a new entry to its gazetteer as enrichment. In this way the system is able to leverage real local and community conversations and outsource this geocoding task to local community members. Hence, new gazetteer entries learned from human reflect local spatial language and local knowledge. 3. Smart footprint recommender rankst he matched candidates by their likelihood of being correct. Based on our understanding of how humans disambiguate multiple interpretations of place names, a computer can play the role of a smart recommender by automating a set of heuristic rules. By presenting the candidate list in the order of likelihood, human annotators are more likely to find the answer by exploring only the top few candidates. This creates savings to human cognitive efforts." 1 and 2 are critical (but with two- how should they be saved/named when hyper specific?), and 3 is gravy. [32]
-Smart recommendation: "prioritize previous annotations", "exploit local gazetteer", "rank search results with regard to a moving focus". See p7 fr more. [32]

2.3.10 Location assignment

During publication
-Geo My WP: Integrates GoogleMaps, Leaflet, and OpenStreetMaps, includes proximity search, point location definition [59]
- CG Geo Plugin: Focused on customer geolocation integration: geotargeting, legal requirements, etc. [60]

2.4 Web applications

2.4.1 Design

-"planning organizations should choose a participation platforms based on the capacities of their organization, the characteristics of the communities that are going to use the tool, user-community norms and rules, and the tool's capabilities." [4]
-"Discussing these factors with planners and decision makers, technologists, and communities who use the tools will add insights on identifying new considerations for implanting online participatory technologies" [4]
-"The geospatial Web 2.0, or Geoweb for short, is a collection of online location-enabled services and infrastructure that engages a wide range of stakeholders in mapping processes." [27]
-Basic components: web participants, user groups, geoblog, blog components, blog profile, blog component profiel, blog description, time points and intervals, blog metrics. Relationships: generalization, composition, collaboration, possession. [27]
-"a general mashup architecture consists of three different participants that are logically and physically disjoint. THe generic architecture has the mashup client sitting on the top and the data sources and services sitting on the bottom. The middle tier is where the

mashup logics reside. It should be noted that the mashup logics for generating mashed content could be either executed on the server or within the web browser. [27]
-Tagging: used GeoBLog with spatial extensions (server side scripting using JavaScript, Visual Studio 2010 2ithin the asp.net code). Requires secure user authentication (login), access to post and transfer messages, and management tools programmes in asp.net. [27]
-Geoportal design: i. spatial database and data server; ii. geospatial web server; iii. Geospatial metadata catalogue server; iv. other catalogue server; v. download tool. [6]
-"deconstructing the user-designer dichotomy... the boundary between user and designer is fluid and configured during the design process, and that users can have multiple identities. In addition to being users, they can perform activities traditionally ascribed to designers by dynamically participating in the ongoing design process." "a user-centered design that deconstructs the traditional power relationship between designers and users through role hybridization by creating an environment where users or designers are able to shift from one role to another, effectively belonging to otherwise two distinc groups." "designers can take the role of users, and users can participate in the design process, with constant interactions within and between the users and designers in order to combine and reconcile design ideas from the wto groups of users and designers." "designer-user interaction models as a combination of interaction of designers, interaction of users and mutual interaction between designers and users with constant communication to increase knowledge sharing, crossing the social boundaries and deconstructing the design-user dichotomy to produce novel design artifacts and increase usefulness." [46]
-"With inter-user communications capability built into the platform (in addition to the periodic face to face meetings and email exchanges), both users and designers were able to articulate their needs, comment on particular linguistic and geographic difficulties, deliberate together and suggest solutions from implementation." [46]
-"In fact, code writers can be considered designers because they create software and algorithms that support evaluations and decisions directly impacting people's lives. The filters that determine the selection and analysis of large and complex data sets on topics ranging from food access to health and nutrition originate from their coders' specific world-views and reflect their implicit bias, inevitably shaping the outcome of research and surveys, which in turn influence policy decisions based on those enquiries." [61]
-"The dramatic rise of data visualization could be traced to hardware factors such as widespread use of high-resolution large desktop displays tied to powerful computers. Other important trends are the increased availability of vast data resources, familiarity with data management software, and innovative web-based software that support rapid display and update of visual information." [62]
-"The datavisualization process (also called 'work flow') starts with the identification of stakeholders and their insight needs. Just as a verbal math problem needs to be reformulated into a numerical math problem, the verbal or textual description of a real-world problem presented by a stakeholder must be operationalized (i.e., reformulated into a data visualization problem so that appropriate datasets, analysis and visualization workflows, and deployment options can be identified)." [18]
-"The task hierarchy provides insights about the 'bigger-picture' of each atomic task - a task that does not contain any subtasks - helping ensure that we always consider higher-level goals." [21]
-"As tasks often have dependency relationships with each other, there can be a required task sequence. E.g., one has to view data before reasoning about it, and an overview visualization needs to be presented before selecting and examining outliers. Therefore, visualization tools should ont simply fulfill discrete tasks without providing for a natural

flow between consecutive steps. Integrating task sequence information can help mingate the user's mental burden and provide a better overall user experience." [21]
-"While many abstract tasks are similar (e.g. discover patterns and trends, reason about outliers), visualization tools often are used in a domain-specific context... We cannot simply peel off the domain-specific context, abstract to common data visualization tasks, and then design for those tasks. Context is key." [21]
-"When designing data visualization tools it is crucial to choose visual and interaction designs that solve the real-world problems of our target users. Task abstraction is widely used to support this process by characterizing user tasks as domain- and interface-agnostic abstract tasks for which design best practices are known." [21]
-THree step hierarchical task abstraction: "(1) understand the problems and data of domain users; (2) perform hierarchical task analysis to understand user tasks, goals, and the relationships among the tasks as part of a larger analysis process; and (3) abstract the user domain tasks." [21]
-"Once constructed, the hierarchical task and data type abstractions should inform visual encoding and interaction design." [21]

2.4.2 Open Source

-"To develop on an open source platform is extremely vital when huge databases are to be created and consulted regularly for region planning at different scales particularly satellite images and maps of their locations". [26]
-OGC Standards -Trailblazers set the tone for future scalable impact and change. Foundational development and openness are key [14]
-"Transparency underpins integrity and legitimacy. Trailblazers have nothing to hide. They are working towards a better future and are transparent as to how their product road map will lead to better solutions over time." [14]
-"Trailblazers share their technological knowledge and intellectual property with other players in the industry. THey recognize that it is more important to increase the pie than to increase only their share of it. From creative commons and open source tools to commercial arrangement, such as commercial licensing or training and support services, trailblazers find ways to push the market with their knowledge. They do this in a way that enables others to follow them towards a higher standard without compromising their competitive position." [14]

2.4.3 Data

Volunteer generated information sources
-VGI and geoweb: "there is mounting evidence that institutions can use VGI as a mechanism to build a local capacity to support collaboration, supplement traditional data sources, and inform decision making." The VGI paradigm "Is enabled by teh GeoWeb, locatinoaware devices, and citizens acting as sensors, and their tools and resources for collecting and processing geographic information from volunteers are readily available." [27]
-"VGI involves the creation and dissemination of geographic data provided voluntarily by individuals and overlaps with PPGIS in that both involve the investigation and identification of locations that are important to individuals." [23]
-"PPGIS projects are often implemented to inform planning and policy issues while VGI systems may have no explicit purpose other than participant enjoyment." [23]

-“Given well-defined insights needs, relevant datasets and other resources can be acquired. Data quality and coverage will strongly impact the quality of results, and much care must be taken to acquire the best dataset with data scales that support subsequent analysis and visualization.” [18]

Data harmoziation -Data harmonization: “Some geoportals not only provide data as they come from the original source, but they are also able to provide spatial data and geo-information coming from different sources into a common (i.e. standard) format... Data harmonization can help to receive, process, and exchange data, and to ensure interoperability, because the harmonized data and information are accessible to end users at their demands.” Federated approach: [6]

-Federated approach: “participants to agree on common specifications in terms of metadata, data models, and service interfaces, typically based on international de-jure or defacto standards.” Brokered approach “leveraged middleware between the client and the server tiers by addressing heterogeneity through mediation (i.e. of metadata and data models) and adaptation (i.e. of interfaces). In the brokered approach, the brokers as key component are dedicated to mediation and haronization. The brokered approach has been particularly appealing for those wanting to build Systems-of-systems, e.g. distributed infrastructures connecting systems, which keep their autonomy at a certain degree.” “Generally, the brokered approach is more effective in addressing variety and heterogeneity when central authority cannot be easily achieved.” [6]

-Basic functionality: “metadata registry, data discovery through a catalogue service, data visualization, and data access.”See section 2.3.4 for more info [6]

-“A geospatial metadata catalogue provides data descriptions in terms of metadata (e.g. contributor, data type, language, contact point, keywords, and dataset identifier for data localization and indexing). In addition, the metadata catalogue is often used for implementing harmonized data discovery. Since users are typically interested in finding datasets matching specific constraints, the data discovery functionality is one of the basic functions that geoportals offer. SPecifically, geoportals providing data discovery generally allow searching datasets along the who, when, where and what axes, that is, by geo-location (where), data provider (who), time range (when), thematic layer, and keywords (what). The user interface provides graphical tools, like a bounding box on a map, to set spatial and temporal constraints. Moreover, users can be directed to a gazetteer, a thesaurus, or other knowledge bases for better scoping their query. Various approaches have been developed to enhance geoportal search capabilities, e.g. the use of thesauri, ontologies, and semantic text matching algorithms”. [6]

-“Big data are posing challenges that are usually referred as V challenges: large Volume, high Velocity, and wide Variety. On the geoportal user interface side, they have an impact on discovering, visualizing, processing, and storing big Earth data. Specific challenges come from teh continuity of data updating, methods of data harmonization, and mulitple functionalities fro professional users.” [6]

-Top down approach: “administrative approach”. Support “incorporate large amounts of spatial datasets and geo-information products, multi layered and multi types of data for earth sciences.” [6]

-Top down approach: “administrative approach”. Support “incorporate large amounts of spatial datasets and geo-information products, multi layered and multi types of data for earth sciences.” [6]

-Top down approach challenge: sustainability and homogeneity, insufficient searchability for nuanced use. Open Geographic Modeling System (OpenGMS) to combat this. [6]

-“The data systems, where the geoportals connect, are leveraging different methods for coordination, in terms of federated coordination (top-down) and broker coordination (bottom.up). Both coordination solutions are able to connect distributed data servers and information systems, and to facilitate complex systems for solving big data storage, management, and assessment problems. The federated approach generally shows better performance, but it requires that participants agree on participating in the overall system bylaw or their own interests. If this is not available, a brokered approach is the only one that can be adopted. In addition, the brokered approach is better, if the requirements cannot be fully defined allowing one to select the best federal agreement. Therefore, a hybrid data system architecture for a complex data system is recommended for meeting the challenges posed by big data.” [6]
-Data integration: “The discipline of data integration comprises the practices, architectural techniques and tools for achieving the consistent access and delivery of data across the spectrum of data subject areas and data structure types in the enterprise to meet the data consumption requirements of all applications and business processes.” (Gartner) [33]
-“Performing temporal inference using disparate applications for each data source is challenging. Despite several advances in data interaction and visualization, device interoperability and data integration remains problematic.” [21]

Data access
-Geoportal design: Spatial metadata catalogue server: “The geospatial metadata catalogue provides metadata in accordance with OGC and ISO standards and is retreived by teh geoportal” [6]
-Geoportal design: other catalogue server “Other catalogues could be supplemented by metadata catalogues for a data registry. A well-known catalogue is theGazetteer Service which can host a location-based feature dataset” [6]

2.4.4 GIS design

-“the design of the event model is composed of two parts: Spatiotemporal and Attribute information. Spatiotemporal information is polymerized by Feature, Temporal and Geometry. Among them, Temporal is composed of the time instant and time interval. Geometry is described by Position coordinate and Coordinate system, which belongs to Point, Line and Polygon. Attribute information is composed of Border type, Feature type, and Rss:Item; the model provides an extension of the elements by Rss:Item.” [27]
-Geoportal design: Spatial database and data server “Spatial databases store and maintain the data that will be delivered. The database could be on the local server or distributed servers.” [6]
-“data storage and management is one of the critical technical components for addressing the volume challenges of big data. Currently, relational databases (e.g. PostgreSQL), NoSQLs (e.g. MongoDB and HBase), and distributed file systems (e.g. Hadoop) are widely used for big data storage and management. These technologies are also important for geoportals, although the technical component for big data storage and management still requires more theoretical and practical developments.” [6]

2.4.5 Web services

-Standards: Geoportals are usually compliant with several OGC standards, wtih some of the most popular ones being the Web Coverage Service (WCS), the Web Processing Service (WPS), the Web Coverage Processing Service (WCPS), the Web Map Service (WMS), the Web Feature Service (WFS), and the Catalog Service for Web (CSW). GeoServer [6]
-Geoportal design: Geospatial web server: "Geospatial web services provide greographic data to the user through forms of map data services uder the guidance of OGC and ISO standards. Open-source options for web services are capable of hosting these web services, e.g. GeoServer and MapServer [6]
Example:

- Twitter map [63]
- COVID visualization [62]: Good design and professional practice is to show the sources of data, name the contributors to the visualization, and provide an email address for comments, corrections, and questions. [62]Example of Johns Hopkins University visualization of COVID data. References of John Snow and WIlliam Playfairs addition to geodisplays to distill information and draw new spatial conclusions [62].Use of spatial data to "persuade, understand current conditions, and predict future outcomes based on behaviors". [62] Many references to additional display options (weather.com, microsoft power bi, esri, visual action, etc. "These widely used COVID-19 visualizations depend on maps, bar charts, line chars, and scrolling lists, however designing them to fit on small mobile-device displays requires skillful programming." [62]

2.4.6 Graphic user interface

-"The interactive graphical user interface (GUI) allows for data visualization manipulation and sharing." [26]
-"The six areas should be covered on the geoportal interface are including metadata, data thumbnail, data description document, data sample, data entity (i.e. download link), and connections and groups. The goal of mode is to provide a more detailed description of the data on the geoportal and enable users enough information for faster data positioning and discovering." Main elements of a geoportal. [6]
-"A well-designed GUI is beneficial for interacting with the user and providing an easy-to-use experience and powerful visualized interactivities." [6]
-"Geoportals tend to prefer adopting an atlas-like user interface (using OpenLayers and Leaflet). THis may result in the requirement of visualizing geographic distributions of the data and of the functionality for visualized data selection. An atlas-like user interface, as the metaphor, is recommended for geoportal design." Validation of openlayers and leaflet. [6]
-"In order to address some Volume, Variety, and Velocity challenges on the user interface side, the next generation geoportals should specify the time-series investigation, graphs and charts for visual analytics, and 3D and 4D visualization widgets through dynamic visualization tools." [6]
-"For an effective presentation, the news must be shown in an informative, aesthetically pleasing manner, but this must not overwhelm the viewer." [42]
-"Because it is inevitable that multiple news stories will mention the same location, we also need a strategy for dealing with occlusion, since we do not want to place markers on top of each other." [42]

-Dynamic map labeling: "placing text labels tied to geographic coordinates on a map, usually requiring that some corner or edge of the label's bounding box touch the coordinates associated with the label." Leverage a precomputed conflict graph ("labels correspond to nodes, and edges exist between labels that overlap"). [42]

-"some occlusion of markers is tolerable, as long as the more significant story markers are placed above less significant markers. One exception to this rule is when markers exactly coincide - that is, when several stories mention the same geographic location. It is unacceptable to place markers at the exact same coordinates on the map, as users cannot infer that many stories refer to that location. This is often a problem with large cities, as they are a part of geographic focus of many news articles." Spiral strategy: "this allows significant geographic locations to have more of their stories visible, at the expense of accuracy in marker placement. However, due to their regular shape, these spirals are usually easy to identify and do not contribute significantly to user confusion."Ideally the inclusion of custom reference points/areas will alleviate this issue, though more popular POIs will likely include several concurrent news stories. [42]

-"Many of the currently popular visualization follow the well-established Information Visualization Mantra: 'Overview first, zoom and filter, then details-on-demand.'" [62]

-Visual Information Seeking Mantra: overview first, zoom and filter, then details-on-demand [17]

-Data types: 1D, 2D, 3D, temporal and multi-dimensional data, tree and network data). Includes detailed descriptions of these. [17]

-Tasks: overview ("gain an overview of the entire collection"), zoom ("zoom in on items of interest"), filter ("filter out uninteresting items"), details-on-demand ("select an item or group and get details when needed"), relate ("view relationships among items"), history ("keep a history of actions to support undo, replay, and progressive refinement"), extract ("allow extraction of sub-collections and of the query parameters"). [17]

-Multi-dimensional: "the interface representation can be 2-dimensional scattergrams with each additional dimension controlled by a slider." [17]

-Multi-dimensional: "the interface representation can be 2-dimensional scattergrams with each additional dimension controlled by a slider." [17]

-Data Scales: "Data variables may have different scales (e.g., qualitative or quantitative), influencing which analyses and visual encoding can be used." [18]

-Visualize "This step can be split into two mian activities: pick reference system (or base map) and design data overlay. The First activity is associated with selecting a visualization type, and the second activity is associated with mapping data records and variables to graphic symbols and graphic variables." [18]

-Deploy "Different interface controls make diverse interactions possible: buttons, menus, and tabs support selection:; sliders and zoom controls let users filter by time, region, or topic: hover and double click help users retrieve details on demand; and multiple coordinated windows are connected via link and brush." [18]

-Interpret: "Finally, the visualization is read and interpreted by its author(s) and/or stakeholder(s). This process includes the translating the visualization results into insights and stories that make a difference in the real-world application." [18]

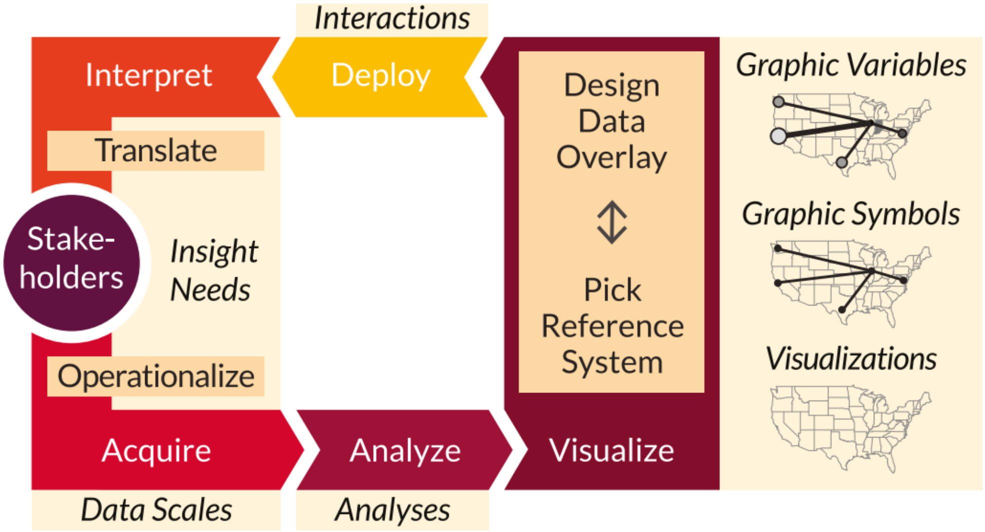


Fig. 6. Process of data visualization construction and interpretation with major steps in white letters. Types identified in Table 1 are given in italics, and an exemplary US reference system and sample data overlays are given in *Right*.

Figure 1: Process of data visualization construction and interpretation with major steps in white letters. Types identified in Table 1 are given in italics, and an exemplary US reference system and sample data overlays are given in Right. (Borner2019)

2.4.7 Data manipuliation

-Geoportals can also offer online processing functionalities, ranging from basic transformations (e.g. coordinate reference system re-projection, sub-setting, format mapping) to complex algorithms for classification, and statistical analysis. That may require an underlying infrastructure providing a workflow engine to orchestrate the multiple actions required to implement the required processing.” [6]
-Spatial queries require indices (such as R-trees or quadtrees) for efficient execution. [44]
-Zoom: “smooth zooming helps users preserve their sense of position and context”. [17]
-Filter: “By allowing users to control the contents of the display, users can quickly focus on their interests by eliminating unwanted items.” [17]
-History: “It is rare that a single user action produces the desired outcome. Information exploration is inherently a process with many steps, so keeping the history of actions and allowing users to retrace their steps is important.” [17]
-“Dynamic queries can reveal global properties as well as assist users in answering specific questions.” [17]
-“dynamic visualizations: overview, zoom, filter, details on demand, relate (viewing relationships among items), history (keeping a log of actions to support undo, replay, and progressive refinement), and extract (access subcollections and query parameters)” [18]
-“Keim distinguishes zoom filter, and link and brush as well as projection and distortion techniques as a means to provide focus and context.” [18]
-“Brehmer and Munzner covers two main abstract visualization tasks. The first is ‘why’, which includes consume (present, discover, enjoy, produce, search (lookup, browse, locate, explore), and query (identify, compare, summarize). The second is ‘how,’ which consists of encode, manipulate (select, navigate, arrange, change, filter, aggregate), and introduce

(annotate, import, derive, record).” [18]
-Other setup: “data and view specification (visualize, filter, sort, derive), view manipulation (select, manage, coordinate, organize), and process and provenance (record, annotate, share, guide).” [18]

2.4.8 Security

-“some people might be concerned about sharing their identify in online environment or making their profile information visible, as they are worried about organizational and social threats”. [4].
-“For implementing these common functionalities, additional functionalities are often required. For example, a geoportal should implement a subset of functionalities concerning privacy and security aspects, e.g. authentication, data access control, logging. Moreover, the geoportal administrators need dedicated management functionalities. [6]

2.4.9 Products

-Recommended services for geoportals: catalogue, preview, access, and online analysis and processing [6]
RSS
-“Publish/Subscribe has emerged as a communication paradigm able to facilitate the development of complex distributed applications in open network environments. THe strong decoupling it introduces between communication parties enables applications to publish information without being aware of the identities of potential receivers or even of their existence. Similarly, it enables receivers to issue subscriptions that express their interests in messages with a given content regardless of the identity of their publishers.” [27]
-“GeoRSS is an emerging standard for RSS feeds to be descried b location or geo-tagged in a standardized way in which the location is encoded. Many map APIs support GeoRSS feeds with either coordinate (lat/long) data or address information specified in XML items. “ [27]
-Geoportal design: Download tool: “A download tool incorporates the capability of obtaining data and metadata. Metadata could be downloaded by XML structured files, and data could be downloaded as raster or bector based data outputs through various ways, such as FTP” [6]
-GeoRSS feed suppor [55]

API
-“it is suggested to provide tools (e.g. API, widgets, configuration) to create community portals, and applications tailored to specific users. THey would allow developing, discovering, and running of applications to support different online and cloud based scenarios in particular for big earth data processing… Geoportals could involve online spatial analysis functionality for addressing geospatial analysis tasks, e.g. online computing environments and geospatial processing web, online analysis, and cloud computing.” [6]
-“Geoportals could also provide APIs to enable developers to create alternative user interfaces to data systems, including community portals and mobile/desktop apps.” Incorporate API functionality. [6]
-“System components communicate using open application programming interfaces (APIs) so that each component can be replaced by a different technology if necessary.” GeoAn-

notator [46]

2.4.10 Dashboard design

-“Viewing multiple folded and aligned units of event sequences simultaneously and in their entirety would require considerable screen space. However, aggregation can obscure rapid changes and outliers. Overview and detailed approaches are beneficial in these cases.” [21]
-“The benefit of superimposing time series from multiple days is to allow direct comparison without changing the view. However, these line charts may cause significant visual clutter and conflicts. Superimposing distinct views can sidestep this problem. [21]
-“Our participants appreciated and were able to effectively use superimposed distinct visualizations of multivariate data in the overview and detail panels.” [21]

2.4.11 Internationalization and localization

-“Localization refers to the adaptation of a product application or document content to meet the language, cultural and other requirements of a specific target market (a ‘locale’). [64]
-Localization: l10n [64]
-“Internationalization is 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.” [64]
-Internationalization: i18n [64]
-“Retrofitting a linguistically- and culturally-centered deliverable for a global market is obviously much more difficult and time-consuming than designing a deliverable with the intent of presenting it globally.” [64]
-“So ideally, internationalization occurs as a fundamental step in the design and development process, rather than as an afterthought that can often involve awkward and expensive re-engineering.” [64]

2.4.12 Examples

Geo my WordPress
-“Geotag any of your post types, buddypress members and other components” [59]
-“Create unlimited advanced, proximity search forms to search and find any of the geotagged components of your site” [59]
-“Add geographic location to any of the registered post types of your site. Display post location on a map, and create proximity search forms to search and find posts based on address, distance categories and more.” [59] -Based on Google Maps API, also supports Leaflet and OpenStreetMaps [59]
-Reviews: 4.6/5 stars. [59]
-Appears to only be for point data (not area) [59]

NewsPack
-A wordpress model for news sources (mostly american) [65]
-The backoffice 60% of global news rooms [65]

-Doesnt' allow for customization [65]Opportunity to create a plugin for wordpress to accomplish this -Easy to work in [65]
-Mostly tailored to US local news [65]
-Doesn't incorporate or recommend a geolocation element [66]

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 [6].

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 [16]. 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 [44]) 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 [47].

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 [23]. 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 [12].

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. [44]. 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) [23]. 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 [37].

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 [67] 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) [68].

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 [69]. 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 [19].

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 ([5,26,42,70]), 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 ([6,26,70]) 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 ([70]).

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 8.

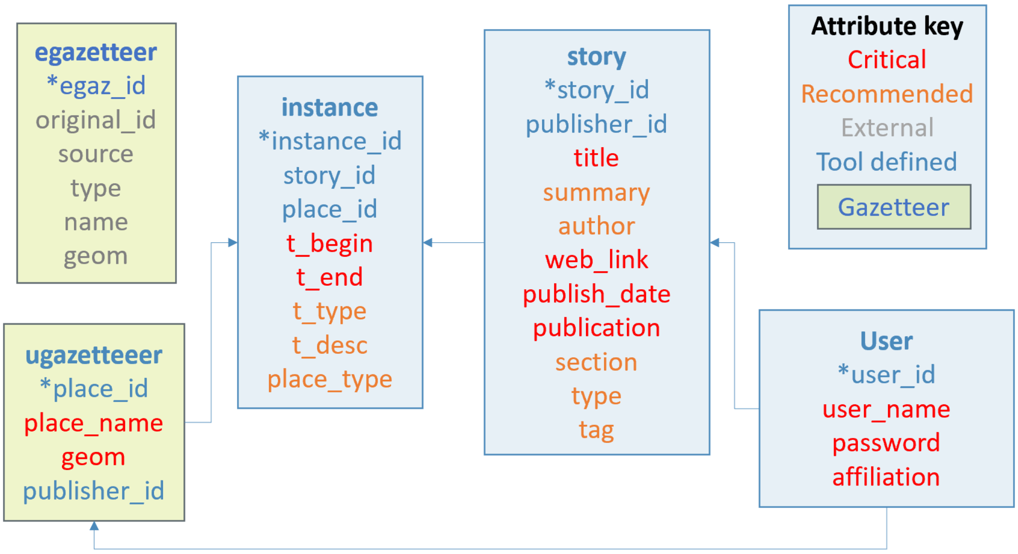


Figure 2: Data model

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 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.

Source	Data	#	Contents' dates	Scale
CML	Boletim 1392	36	Mar - Oct20	address - municipality
	Notícias	42	Oct20	sub-bairro
JFC	Notícias Newsletter	21	Sep - Dec20	sub-bairo - freguesia
JCE	Agenda	4	Jun-Dec20	freguesia
	Notícias	10	Oct20-Jan21	address - freguesia
Público*	ípsilon impar Público Fugas p3 (uncategorized)	83	Oct20-2030	address - inter-municipality
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 direclty 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. Priority 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. Other fields are helpful but not critical for inclusion.

Attribute	**Type**	**Priority**	**Visibility**	**Note**
Title	Text	Critical	Yes	
Summary	Text	Low	Maybe	Keyword search priority
Contents	Text	High	No	Keyword search
Photo	Web link	Low	Maybe	
Section	Text	High	Yes	Thematic filtering
Themes	Text	High	Yes	Thematic filtering
Times	Text list	High	Yes	Temporal filtering
Places	Text list	High	Yes	Geospatial filtering
Referenced articles	Web link	Low	Maybe	Suggestion
Related articles	Web link	low	Maybe	Suggestion
Author	Text	High	Yes	
Source	Text	Critical	Yes	
Publication date	Date	Critical	Yes	Default temporal filtering and extraction
Link	Web link	Critcal	Yes	

Table 2: Corpora attribute collection

4.5.2 Basemaps

Source	Name	# Records	Geometry	SRS
CML Geodados	Quarteirões	1086	Area	EPSG:4326
	Grandes parques e jardins de lisboa	190	Area	EPSG:4326
	Rede viária	3763	Line	EPSG:4326
	Limite do conselho	1	Area	EPSG:4326
	Rede ferroviária subterrânea	1	Line	EPSG:4326
DGTerritório	CAOP 2019 - Continental	3,223	Area	EPSG:3763
	Bairros			
	Address			
Total		**8,224**		

Table 3: Base map data

4.5.3 Gazetteers

-Include feature codes from GeoNames to help with disambiguation of human annotators selecting toponyms [46]

Source	Name	# Records	Geometry	SRS
GeoNames	PT	37327	Point	WGS84
OSM PT	Waterways	49,376	Line	WGS84
		54,041	Area	WGS84
	Transport	18,696	Point	WGS84
		1,078	Area	WGS84
	Traffic	84,756	Point	WGS84
		35,339	Area	WGS84
	Roads	1,035,765	Line	WGS84
	Railways	8,310	Line	WGS84
	POIs	100,214	Point	WGS84
		71,335	Area	WGS84
	POWs	2,133	Point	WGS84
		7,358	Area	WGS84
	Places	23,594	Point	WGS84
		1,971	Area	WGS84
	Natural	89,483	Point	WGS84
		1,369	Area	WGS84
	Landuse	209,330	Area	WGS84
	Buildings	1,060,745	Area	WGS84
Total		**2,892,220**		

Table 4: Gazetteer data

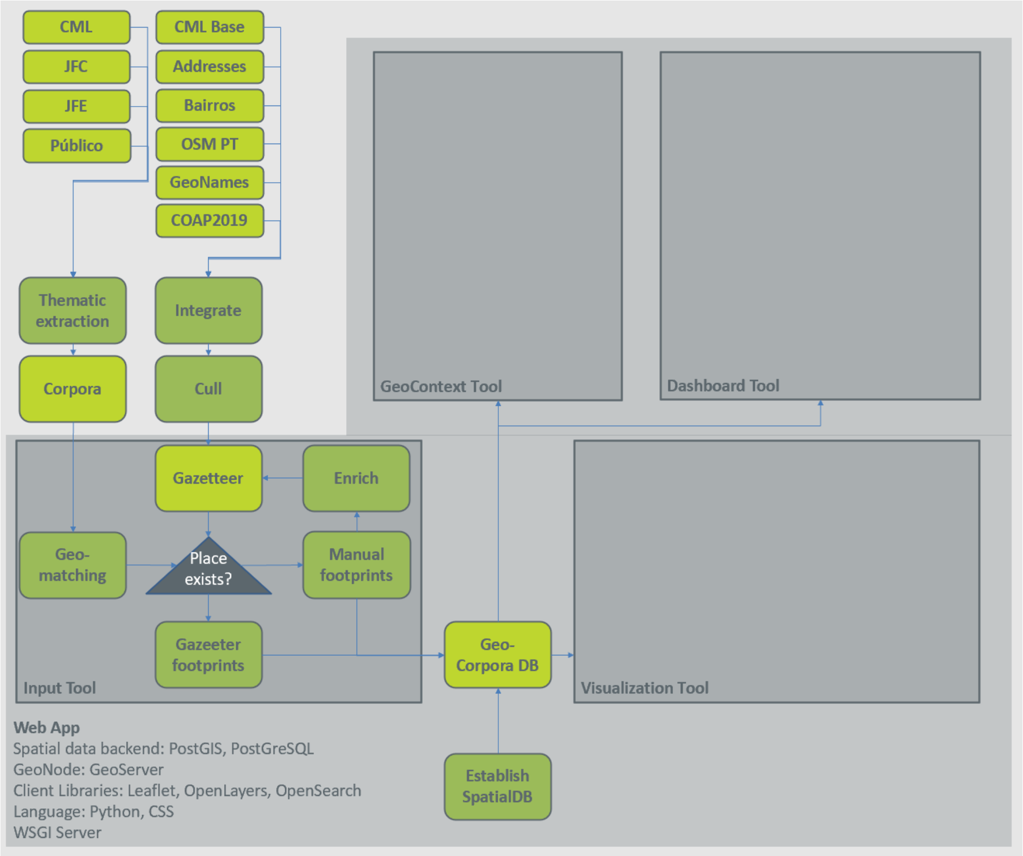


Figure 3: Preliminary methodology

4.6 Preprocessing

Load all gazetteer data into QGIS. Establish georeference (note in data breakdown), transform as necessary Manually code all article data. Using multiple sources to ensure continuity across local data sources.

4.7 Initialization of development environment

Selection and integration of resources and tools:

- Database: PostgreSQL

Establish geodatabase structure, accommodating multiple language options, load gazetteer(s) and relevant administrative boundary data Develop and test Input tool: wordpress plugin? Develop and test search tool Develop and test Context tool Translate web app content to portuguese and load translations Migrate site to the server Test system Compare results against mined location results Document results

Future development: dashboard tool

4.8 Tool design

-Define design requirements and views to achieve them [21]

4.9 Testing

- Test in different browsers for feature functionality. [62]

4.10 Validation

1. GDELT: compare Lisbon Oct 2020 to results 2. Automatic extraction over sample corpora to compare results leverage Rivera2020 study - spaCy -"Large scale infomration extraction tasks" [71]
-Free online classes to learn [71]
-Utilize NER to extract exact and complete places. As a baseline comparison to the input tool? Used together? Feed through the system with a "check point" (verify the identified place before input)? [41]
-Use local datasets (Público, CML, Freguesia) from 1 month of each to build and test a backoffice plugin for defining localization of news articles. Ask MAGG or similar to apply the plug in for upcoming articles for a 1 month period and test against the same dataset with NewsStand. Is it richer/more accurate? Is the journalist satisfied with the result? Is the reader? Load the same local base gazetteer (OSM Portugal and Global) [44]
-Use method for extraction on input articles to extract and geolocate place for input comparison to automated methods?? Python and Restful API

5 Results

6 Analysis

6.1 Future opportunities

6.1.1 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.

6.1.2 Application to GeoIntelligence

-Potential future use of this toolset: GeoINT source. [34]
-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. [42]

6.1.3 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." [46]

-“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.” [8]

6.1.4 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. [48]

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.” [42]
-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 [42]

7 Conclusion

8 Future Opportunities

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

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

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

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

[6] 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.

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

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

[9] 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.

[10] 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.

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

[12] 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.

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

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

[15] Giorgia lupi.

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

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

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

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

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] Maryam Bahojb Imani, Latifur Khan, and Bhavani Thuraisingham. Where did the political news event happen? primary focus location extraction in different languages. *Proceedings - 2019 IEEE 5th International Conference on Collaboration and Internet Computing, CIC 2019*, pages 61–70, 2019.

[31] Agnieszka Leszczynski. Platform affects of geolocation. *Geoforum*, 107:207–215, 2019.

[32] 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.

[33] Dean Hintz and Craig Hantke. How government agencies are integrating & delivering data for emergency response, 2020.

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

[35] National geospatial-intelligence agency, 2020.

[36] Artur Jorge Abreu Varanda. Project "oraculo": Extracting events from news streams and mining their spatiotemporal patterns to support un operations in the central african republic, 2020.

[37] 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.

[38] Sabine Witschas. Cross-border mapping-geodata and geonames. Borders in a new Europe, 2004.

[39] Markus Eisl. Searching european data, 2020.

[40] James. An attempt to extract geo-location from text, 2020.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

[47] 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.

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

[49] ESRI. Adjust how locations and attributes are extracted.

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

[51] Openstreetmap.

[52] Nga: Gns home.

[53] Jacob Granger. Isabelle roughol of borderline podcast, on the pros and cons of 'indie journalism', 11 2020.

[54] Jacob Granger. "mass customisation" of neighborhood data can help hyperlocal news become more sustainable, 9 2020.

[55] Sarah Marshall. #tip of the day for journalists: Geotag your content using a wordpress plugin, 10 2012.

[56] Crosstown: About us, 2020.

[57] Crime, 2020.

[58] Inyourarea, 2020.

[59] Eyal Fitoussi. Geo my wordpress.

[60] Infinitum Form. Cf geo plugin.

[61] Fabio Parasecoli and Mateusz Halawa. Rethinking the global table, 2019.

[62] Ben Shneiderman. Data visualization's breakthrough moment in the covid-19 crisis, 4 2020.

[63] Eric Fisher. Making the most detailed tweet map ever, 12 2014.

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

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

[66] Newspack, 2020.

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

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

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

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

[71] Exposion AI. spacy, 2020.

[72] Jerry Low. How to host your own website: Complete beginner guide, 9 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’.” [42]

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! [22]

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 [72]

Contributor: define!

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

Data visualization literacy (DVL): define [18]

Endonym: “a locally used toponym” [50]

Even-occurring: “all locations where events occurred regardless of whether the event is the event of interest.” [47]

Even-relevant: “all locations where events occurred regardless of whether the event is the event of interest.” [47]

Exonym: “a externally used toponym” [50]

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." [43]

Geographic information retreival: [43]

Geographic name ambiguity: "a given name might refer to any of several geographic locations." [42]

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 [41]; "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." [45]

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." [43]; "the process of recognizing place names in text ('toponym recognition') and resolving them to their coordinates or gazetteer entry ('toponym resolution')" [48]; "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. [46]

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" [6]

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)." [44]

Ground truth: hand coded set of actual locations for training and verification [47]

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." [64]

Localization (l10n): "the adaptation of a product application or document content to meet the language, cltural and other requirements of a specific target market" [64]

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." [42]; " 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 [24]

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 [4]

Ontology: "a formal, explicit specification of a shared conceptualization" [27]

Place: define

Place attachment (PA): define [22]

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. [22]

Reader's spatial lexicon: "those locations that the reader can identify and place on the map without any evidence" [44]

Really simple syndication (RSS):define

Scope: define

Content scope:"the story content's geography" [42]

Provider scope: "the publisher's geographic location" [42]

Serving scope: "based on the reader's location" [42]

Sense of place (SOP): define

Smart city: define

Social capital (SC): define [22]

Spatial data infrastructure (SDI): define [3]

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 [44]

Type by task taxonomy (TTT): [17]

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

Volunteered geographic information (VGI): Define [3]

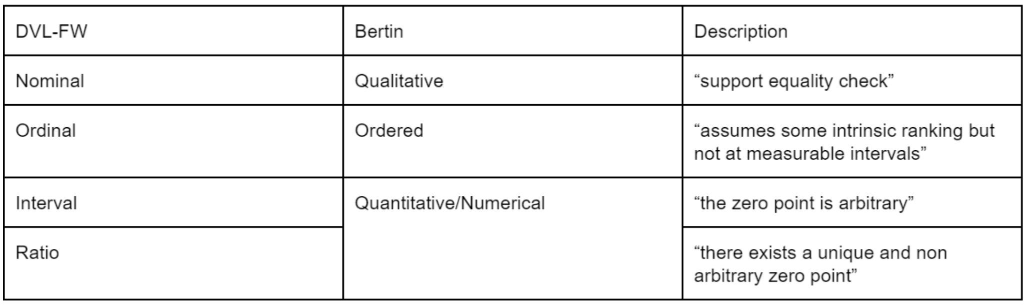
Web application (Web App): a program running on a web server that is accessible via a web browser with internet connectivity.

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

B Data visualization

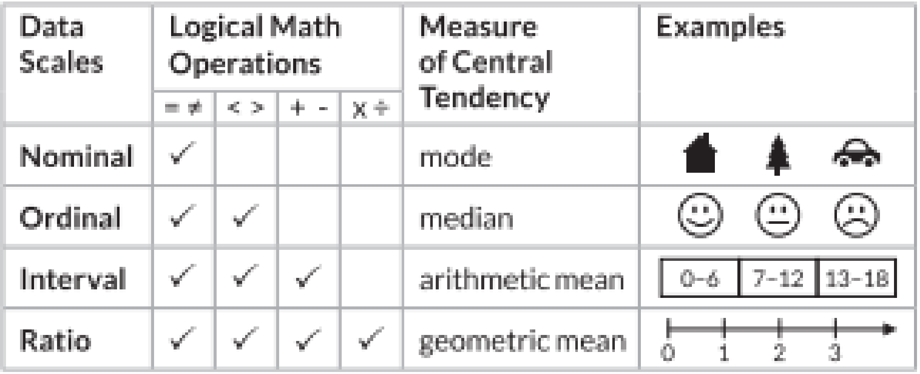
-“quantitative data can be converted into qualitative data (e.g., one may use thresholds to convert interval data into ordinal data). Ordinal rankings can be converted to yes/no categorical decisions (e.g., to make funding decisions). The reverse is possivel as well [e.g., multidimensional scaling converts ordinal into ratio data]” [18]

-“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]” [18]



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 4: 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 5: Logical mathematical operations permissible, measure of central tendency, and examples for different data scale types (Borner2019)

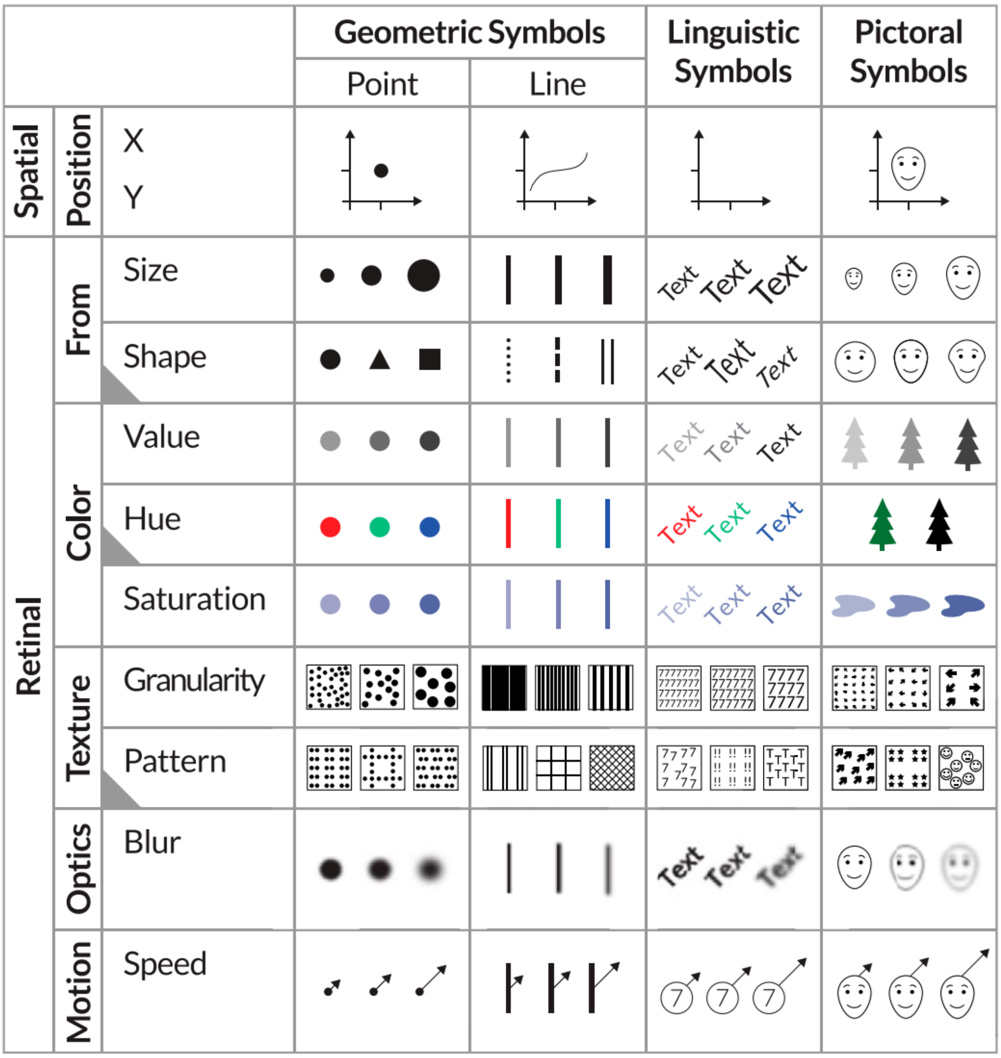


Fig. 3. Four graphic symbols and 11 graphic variables from full 11 graphic symbols by 24 graphic variables set in ref. 34. Qualitative nominal variables (shape, color hue, and pattern) have a gray mark.

Figure 6: 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.” [21]

-“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.” [21]

-“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.” [21]

C Preliminary specification

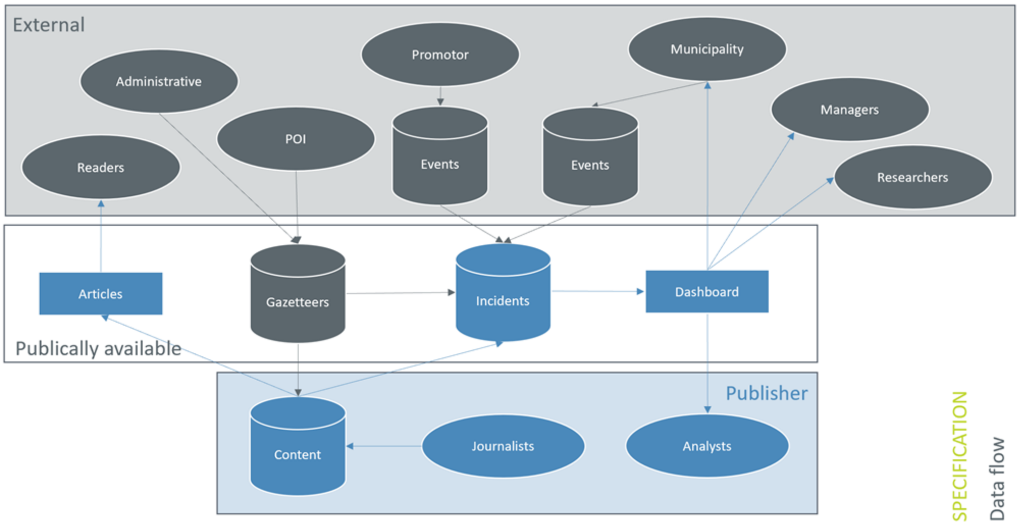


Figure 7: Preliminary data and information flow

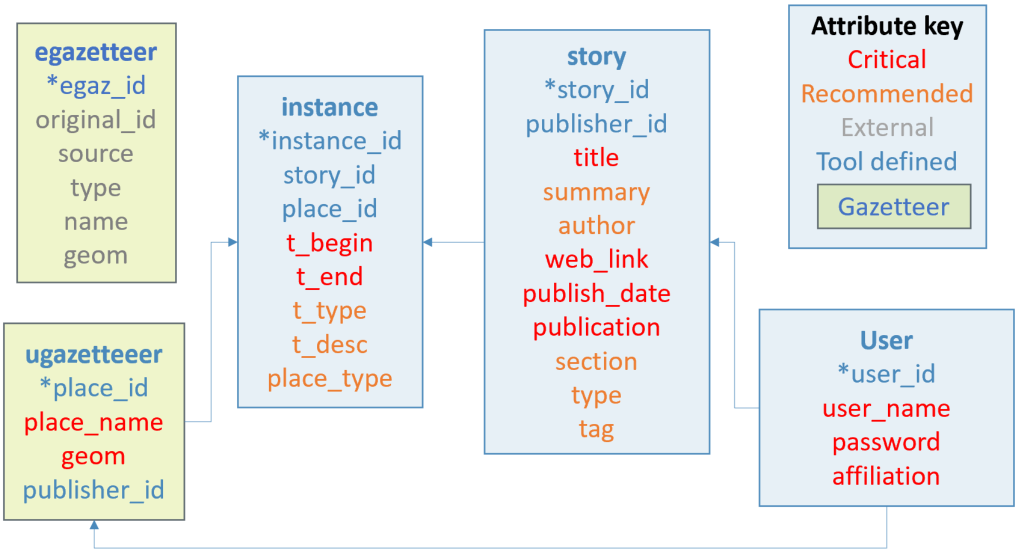
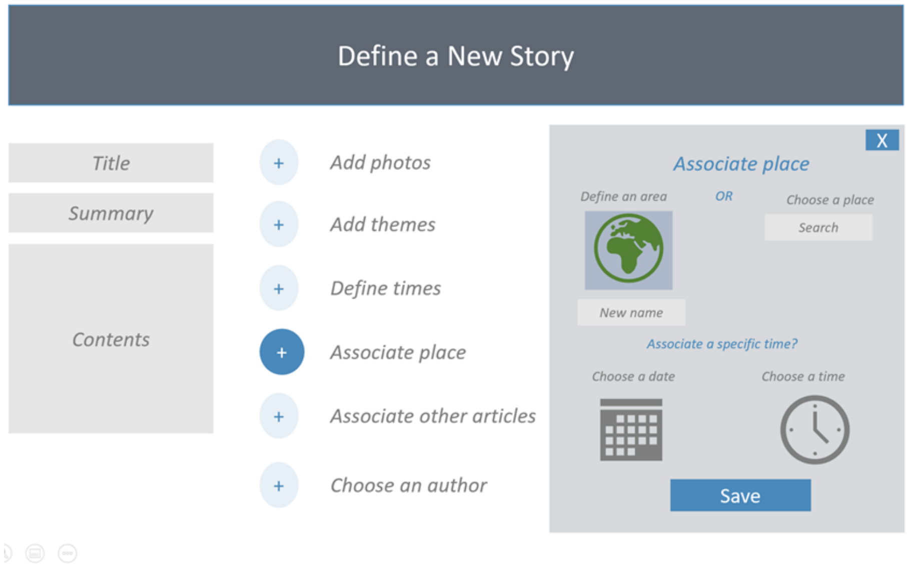
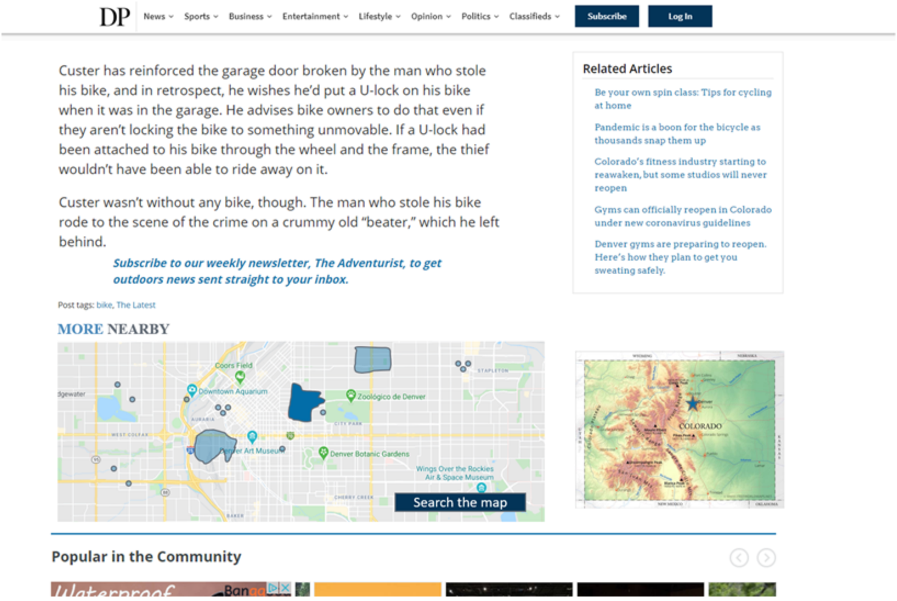


Figure 8: Preliminary data model fo the spatial database



SPECIFICATION
Input UI

Figure 9: Preliminary *Input* layout



Custer has reinforced the garage door broken by the man who stole his bike, and in retrospect, he wishes he'd put a U-lock on his bike when it was in the garage. He advises bike owners to do that even if they aren't locking the bike to something unmovable. If a U-lock had been attached to his bike through the wheel and the frame, the thief wouldn't have been able to ride away on it.

Custer wasn't without any bike, though. The man who stole his bike rode to the scene of the crime on a crummy old "beater," which he left behind.

Subscribe to our weekly newsletter, The Adventurist, to get outdoors news sent straight to your inbox.

Post tags: bike, The Latest

Related Articles

- Be your own spin class: Tips for cycling at home
- Pandemic is a boon for the bicycle as thousands snap them up
- Colorado's fitness industry starting to reawaken, but some studios will never reopen
- Gyms can officially reopen in Colorado under new coronavirus guidelines
- Denver gyms are preparing to reopen. Here's how they plan to get you sweating safely.

SPECIFICATIONS
Context Layout

Figure 10: Preliminary *Context* layout

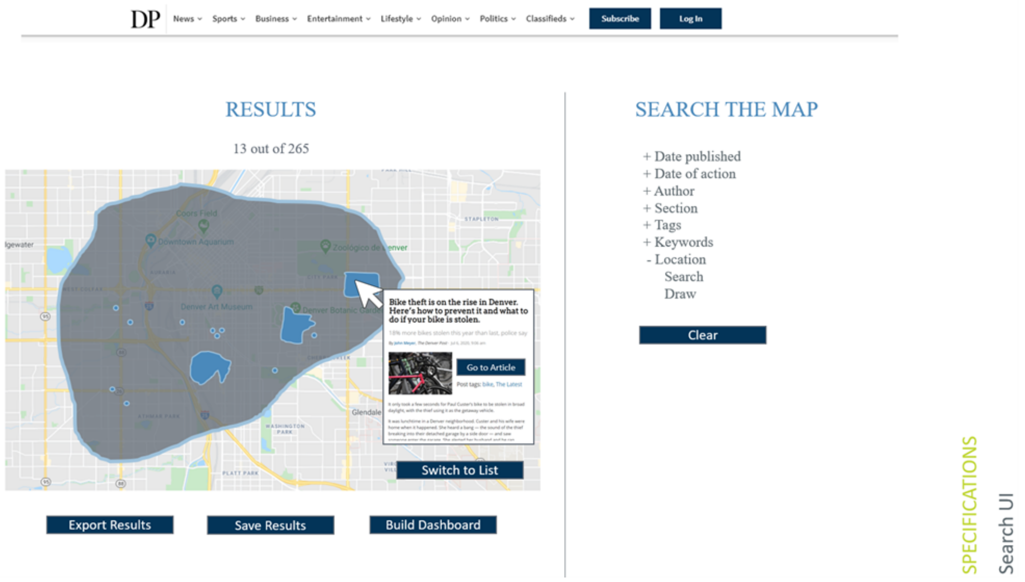
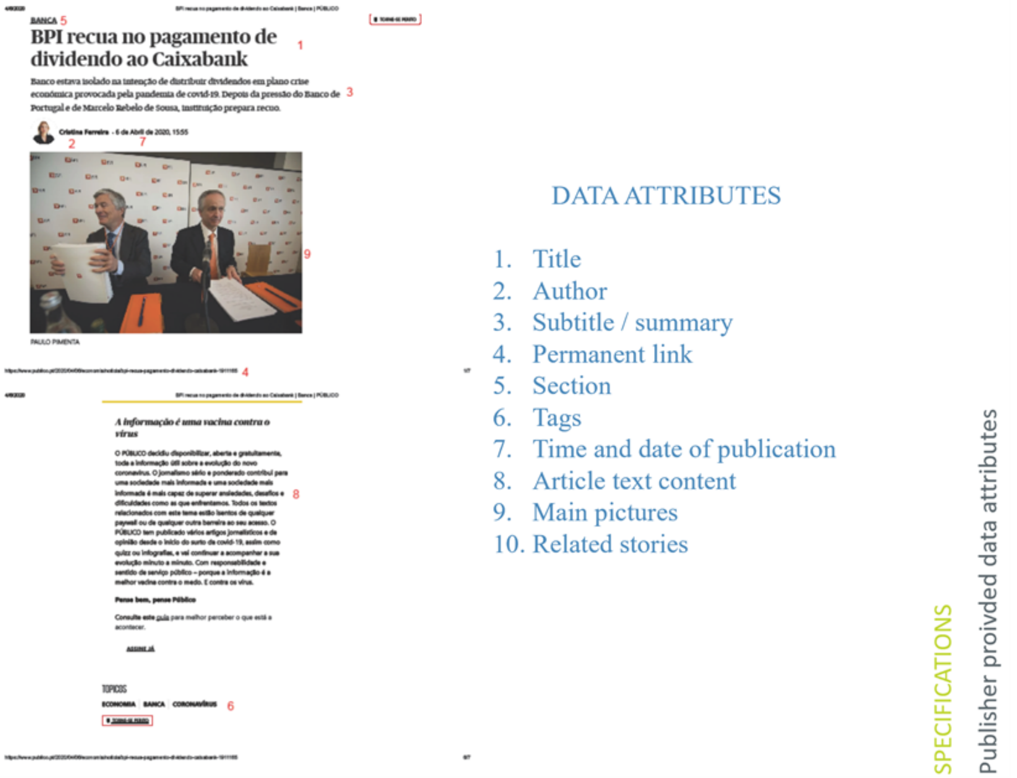


Figure 11: Preliminary *Search* layout



BANCA 5

BPI recua no pagamento de dividendo ao Caixabank 1

Banco estava isolado na intenção de distribuir dividendos em plano crise económica provocada pela pandemia de covid-19. Depois da pressão do Banco de Portugal e de Marcelo Rebelo de Sousa, instituição prepara recuo. 3

Cristina Ferreira 2 · 6 de Abril de 2020, 15:55 7

9

PAULO PIMENTA

4

A informação é uma vacina contra o vírus

O PÚBLICO decidiu disponibilizar, aberta e gratuitamente, toda a informação útil sobre a evolução do novo coronavírus. O jornalismo sério e ponderado contribui para uma sociedade mais informada e uma sociedade mais informada é mais capaz de superar ansiedades, desafios e dificuldades como as que enfrentamos. Todos os textos relacionados com este tema estão isentos de qualquer paywall ou de qualquer outra barreira ao seu acesso. O PÚBLICO tem publicado vários artigos jornalísticos e de opinião desde o início do surto da covid-19, assim como quizz ou infografias, e vai continuar a acompanhar a sua evolução minuto a minuto. Com responsabilidade e sentido de serviço público – porque a informação é a melhor vacina contra o medo. E contra os vírus. 8

Pense bem, pense Público

Consulte este guia para melhor perceber o que está a acontecer.

ASSINE JÁ

TÓPICOS

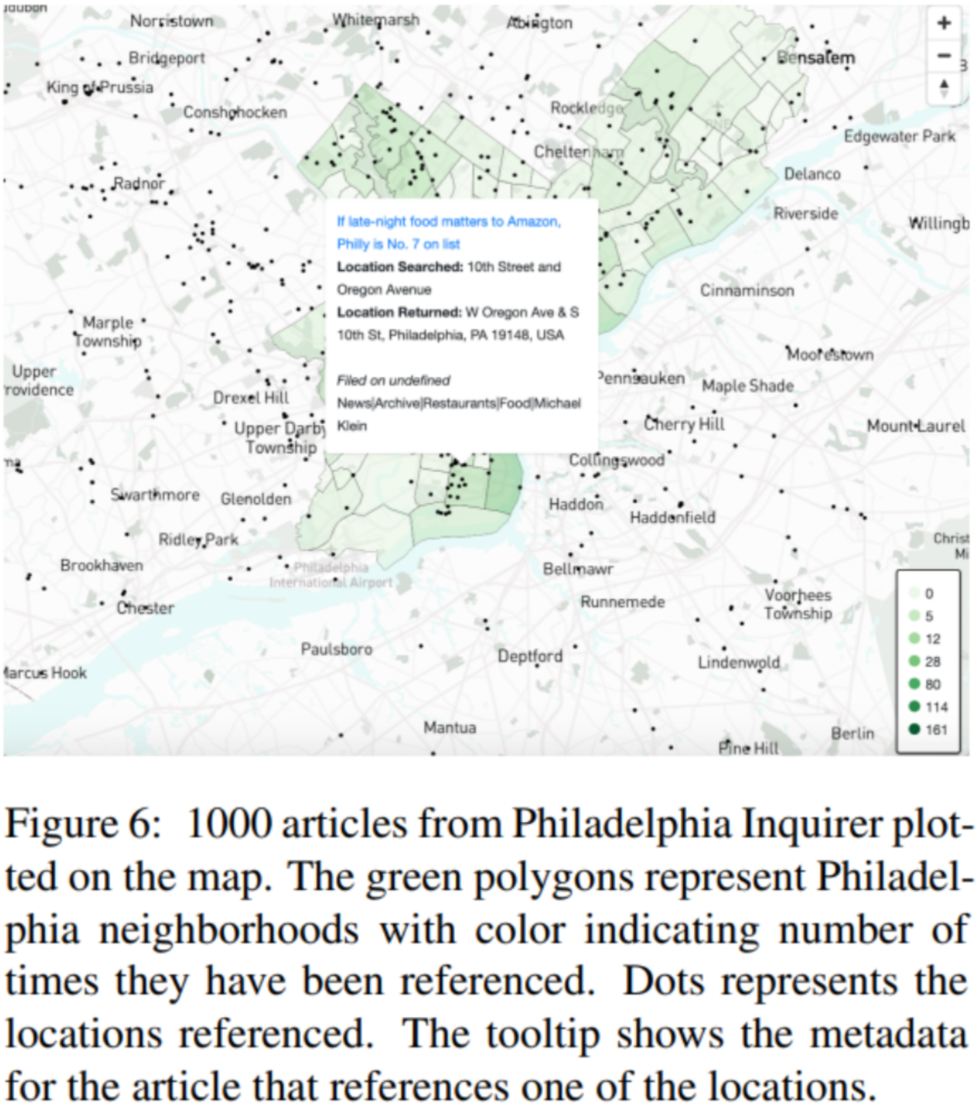
ECONOMIA | BANCA | CORONAVÍRUS 6

DATA ATTRIBUTES

1. Title
2. Author
3. Subtitle / summary
4. Permanent link
5. Section
6. Tags
7. Time and date of publication
8. Article text content
9. Main pictures
10. Related stories

Figure 12: Publisher provided data attributes

D Example GUIs



If late-night food matters to Amazon, Philly is No. 7 on list

Location Searched: 10th Street and Oregon Avenue

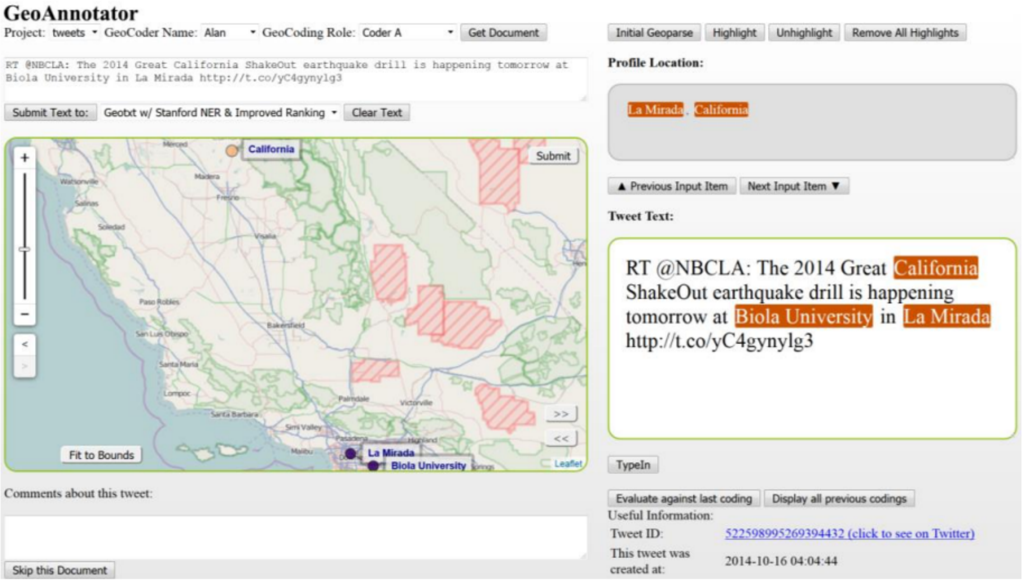
Location Returned: W Oregon Ave & S 10th St, Philadelphia, PA 19148, USA

Filed on undefined

News|Archive|Restaurants|Food|Michael Klein

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 13: Mapped article layout (points) from Gupta2020



GeoAnnotator

Project: tweets GeoCoder Name: Alan GeoCoding Role: Coder A

RT @NBCLA: The 2014 Great California ShakeOut earthquake drill is happening tomorrow at Biola University in La Mirada http://t.co/yC4gynylg3

Submit Text to: Geotxt w/ Stanford NER & Improved Ranking

Profile Location: La Mirada, California

Tweet Text:

RT @NBCLA: The 2014 Great California ShakeOut earthquake drill is happening tomorrow at Biola University in La Mirada http://t.co/yC4gynylg3

Comments about this tweet:

Useful Information:

Tweet ID: 522598995269394432 (click to see on Twitter)

This tweet was created at: 2014-10-16 04:04:44

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 14: Annotation via GeoAnnotator from Karimzadeh2019

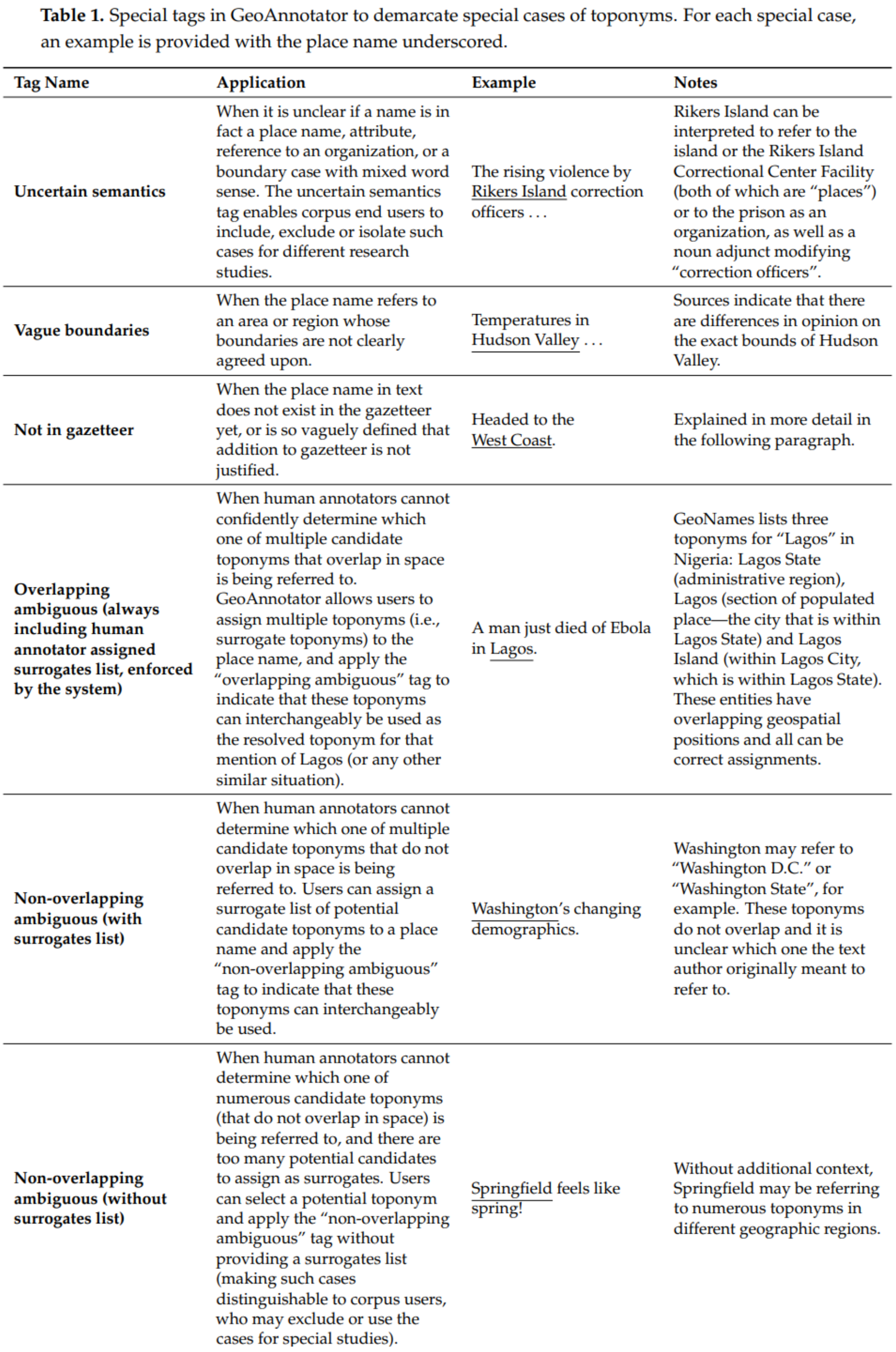
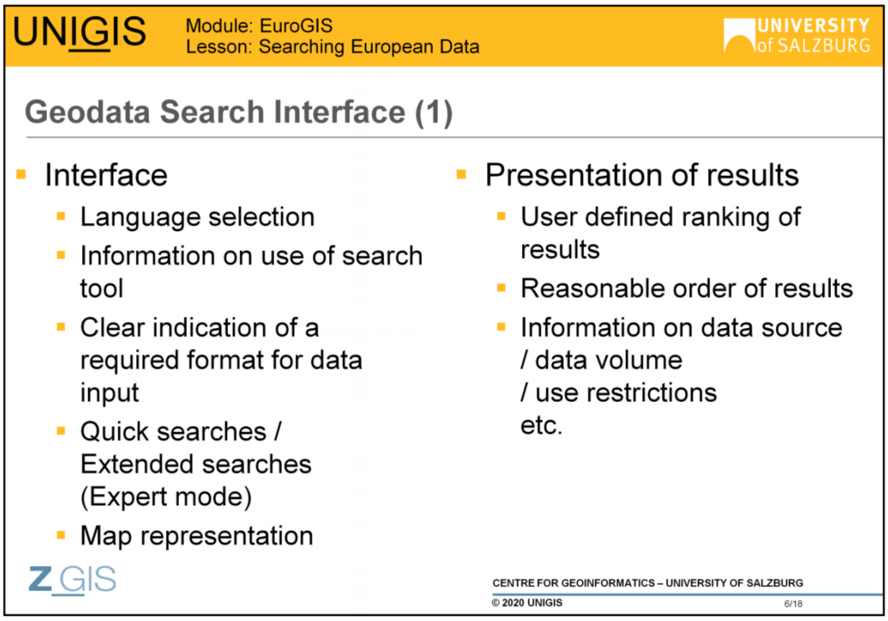


Table 1. Special tags in GeoAnnotator to demarcate special cases of toponyms. For each special case, an example is provided with the place name underscored.

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 15: Annotation ambiguities from Karimzadeh2019



UNIGIS Module: EuroGIS
Lesson: Searching European Data
UNIVERSITY of SALZBURG

Geodata Search Interface (1)

- Interface
  - Language selection
  - Information on use of search tool
  - Clear indication of a required format for data input
  - Quick searches / Extended searches (Expert mode)
  - Map representation
- Presentation of results
  - User defined ranking of results
  - Reasonable order of results
  - Information on data source / data volume / use restrictions etc.

Z_GIS

CENTRE FOR GEOINFORMATICS – UNIVERSITY OF SALZBURG
© 2020 UNIGIS 6/18

Figure 16: Geodata Search Interface from Eisl2020

E Example Architectures

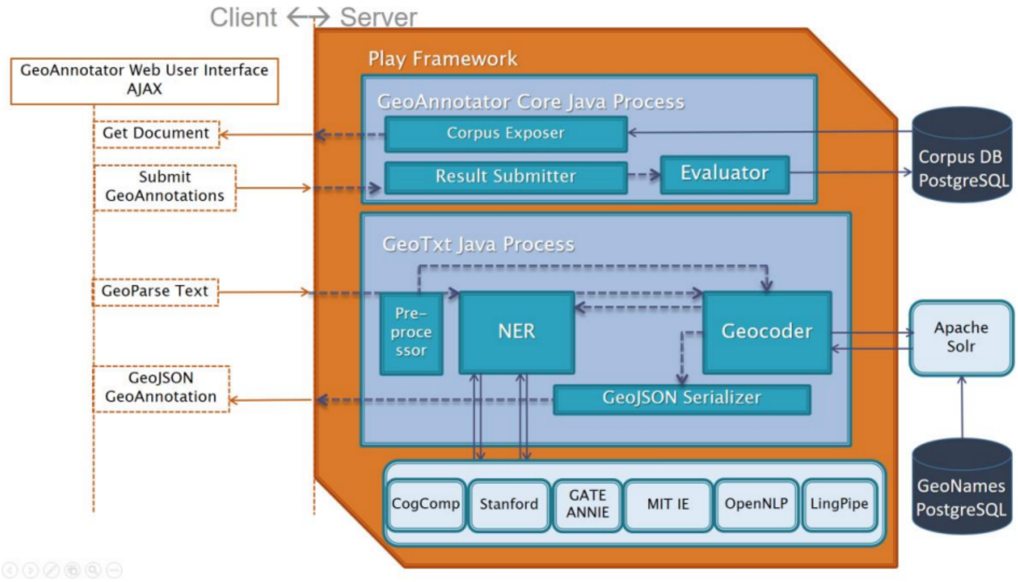


Figure 15. General architecture of GeoAnnotator. The Play framework is used to expose Java application programming interfaces (APIs) as HTTP web-service endpoints.

Figure 17: 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 [46]see section 4.2 for architecture description and inspiration

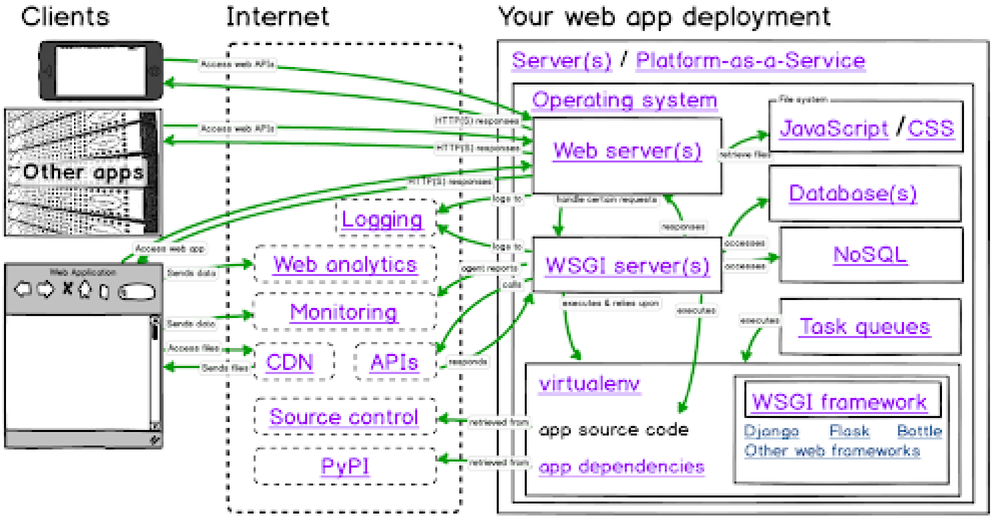


Figure 18: Full stack deployment map from Makai2020

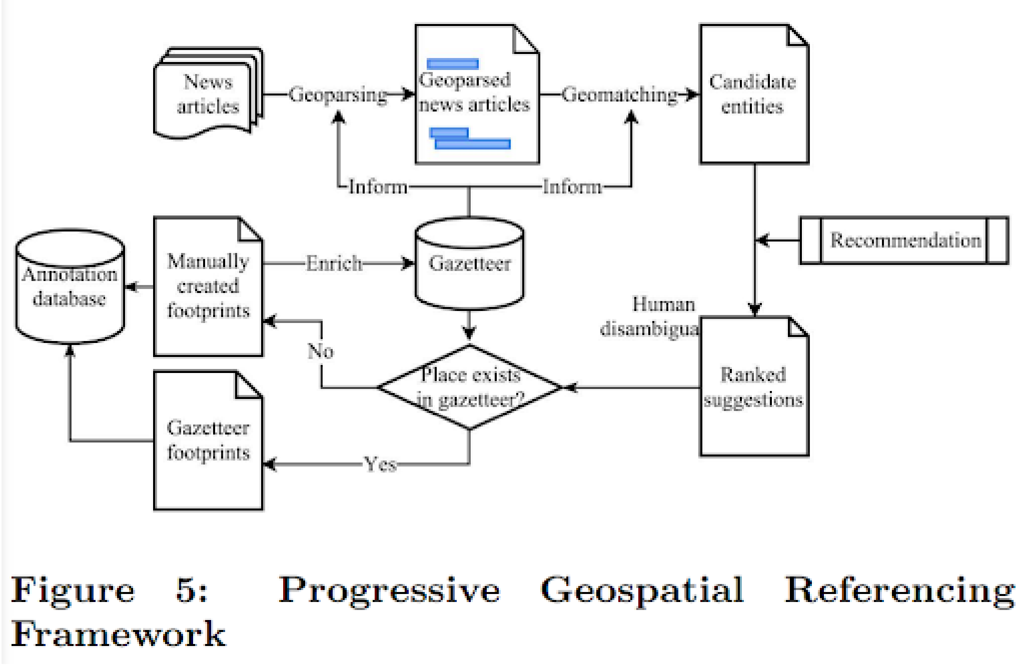


Figure 19: Progressive georeferencing frameowrk from Cai2016

F User stories

F.1 Micronews Portal

[65] - 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