PROPOSAL

Date Submitted: *7/13/2020*

I. Brief Information

Organization Name: Caroline Gilman Wentling (Callie), individual
Project Title: Leak Locale: an open source spatial news web app development project
Project Duration: start date – September 15, 2020 end date – September 15, 2021

II. Project Summary **See list of relevant terms defined in Appendix A.*

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. 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.

III. Problem Statement

The ongoing COVID pandemic has highlighted the value of the visualization of information on a map, for not just specialists to monitor and predict viral outbreaks, but to arm the public with empowering information as well. The value of geographic information systems (GIS) is not just limited to public health services but has already nestled itself into our everyday activities in the form of daily tasks such as navigation and service selection. Applications such as Google Maps, AirBnB, and UberEats allow non-technical users the option to visualize and filter the distribution of various services through spatial (SA), temporal (TA), and thematic attributes (ThA). In the case of AirBnB, a user 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 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 notorious crimes) 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. As such, there is operational as well as academic value in better understanding the spatial distribution of events within a community, such that additional informative insights can be drawn.

IV. Project Goals and Key Milestones

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

1) A spatial database of incidents that supports the association of spatial, temporal, and thematic attributes. *See Appendix C for preliminary data model.*
2) A POC *Input* tool for publishers that allows users to define the place(s) (via search for existing administrative boundaries and points of interest [POIs] through existing gazetteers or definition of new polygons or points via drawing) as well as time of occurrence of incidents. It shall also, of course, preserve or potentially improve upon the association of traditional thematic attributes and keyword search. *See Appendix D for preliminary Input layout.*
3) A POC *Context* map (visualization of an incident on a local map) for integration into each article page. *See Appendix E for preliminary Context layout.*
4) A POC *Search* tool for researchers that allows users to filter by spatial (one or multiple defined places or via drawn definition of the study area), temporal, and or thematic attributes. The results should be displayable via both map and list views, as well as support CSV export functionality. *See Appendix F for preliminary Search layout.*
5) A POC *Dashboard* tool for monitors (publisher, city officials, etc.) to monitor the spatial/temporal development of incidents according to their settings.

The project will use two concurrent "study areas" (news story data from at least one section of a publication for a defined time interval), one in Portugal (such as "Local" in Público for Q3 of 2020) and one in the USA (such as "Colorado News" from The Denver Post for the same time period). In the case that opportunities to include additional study areas such as other sections of publications or additional sources of incident data (such as information from other newspapers or municipalities) arise, these may be accommodated as well, time allowing.

The Web App should support the definition of use in English and Portuguese (leveraging a platform for expansion to other languages) for all elements of the user interface (such as project description, instructions, filters, units, etc.). All data incorporated from external sources (such as news article contents, publisher tags, gazetteer names, etc.) may remain in

their original forms/languages (though alternate forms will be supported if provisioned by the original source).

The project results will be licensed as free and open source such that these can be accessible and leveraged by other individuals or organizations for further development or related projects. Wherever possible, the project will leverage existing open source tools, platforms, and data. However, agreements with data providers may require restriction from public access of their proprietary data.

Note: The project proposed here is not one of automatic place extraction from existing news stories. See Appendix H for more details.

By including study areas in the USA and Portugal, the project seeks to accommodate the culture and business processes of both countries, providing a platform that is useful and valuable to users (whether citizens, officials, or publishers) in both communities. Likewise, the language options of English and Portuguese further support the use and cross-investigation of both countries.

V. Project Methods, Design, and Activities

To support the identified objectives, the following must occur (see section X for major timeline definitions, some elements are already ongoing or complete):

1) Perform literature review of prior art and study of existing relevant platforms and tools.
2) Conduct interviews with stakeholders (publishers, journalists, readers, researchers, and city officials) to establish and prioritize functionality elements.
3) Finalize specifications, mockups of Web App functionality, data model, and finalization of relevant tools and libraries.
4) Initialize the development environment.
5) Receive data from collaborating journals within the defined study areas.
6) Establish incident database, accommodating multiple language options. Load gazetteer(s) and relevant administrative boundary data.
7) Develop and test *Input* tool.
8) Develop and test *Search* tool.
9) Develop and test *Dashboard* tool.
10) Develop and test *Context* tool.
11) Translate Web App content to Portuguese and load translations.
12) Migrate site to the server.
13) Test among stakeholders.
14) Compare against mined location results.
15) Compare against existing media search options.
16) Document results and plan future development.

VI. Estimated Impact and Number of People Reached.

a) 2+ news publication organizations (at least one in USA and one in Portugal)
b) 1 webapp, freely and openly accessible, available in English and Portuguese languages
c) 1 webapp development code, open licensed for further or related future development by any individual or organization.

VII. Overview of the Organization.

This project is being undertaken as an individual. No previous grants from the U.S. Embassy and/or U.S. government agencies have been sought.

VIII. Key Personnel

Caroline Gilman Wentling (Callie), developer, USA. Callie has a bachelor's degree in Electrical Engineering (2012) from the University of Colorado in Boulder. She worked for 5 years in business development and project management at Boulder Engineering Studio, a prototyping and product design firm in Boulder, CO. There, she engaged project stakeholders, specified products and designed project development plans, and managed the projects from early design through production. More recently, she completed a post-graduation program in Smart Cities at NOVA IMS and is currently completing her Masters degree in Geographic Information Systems (GIS) at the same university (including courses in both English and Portuguese languages). At NOVA IMS, she has specifically designed her experience to support the technical design of projects with community improvement and sustainability impacts (see attached CV for additional experience and Appendix I for relevant coursework).

Callie will be responsible for all aspects of the project, including research, development, stakeholder relationship management, testing, and documentation. Advisors include mentors from the University as well as industry relationships. Additional details are available upon request. Existing tools will be leveraged as appropriate into the project to reduce the scope and complexity. Additional services (such as local language experts as translators) will be engaged and are incorporated in the provided budget. Pending full funding of the project, this will be the primary focus of Callie's time and effort.

IX. Project Monitoring and Evaluation

Throughout the project, regular quarterly reports of progress will be generated that describe the major milestones achieved and challenges encountered. Additional support will be sought as necessary to keep to the proposed schedule. Additionally, this project is a development task related to a thesis project of the masters in geographic information systems program at Universidade NOVA de Lisboa Information Management School (NOVA IMS) in Lisbon, Portugal. The project is due in November of 2021, which provides substantial motivation to achieve these objectives on time.

X. Project Timeline of Milestone Deliverables

a) July 2020: finalization of detailed project specifications and data sources, meeting with mentorship team – Lisbon
b) August 2020: full literature review - USA
c) September 2020: receipt of data from study area publishers; engagement of secondary resources - Lisbon
d) October 2020: establish webapp structure - Lisbon
e) January 2020: Functional *Input* tool - Lisbon
f) March 2020: Functional *Search* tool - Lisbon
g) May 2020: Functional *Dashboard* tool - Lisbon
h) June 2020: Full Web App testing; content translation - Lisbon
i) August 2020: demonstration and documentation - Lisbon

XI. Proposed Budget

Type	**Expense**	**Cost**	**Source**
Personnel	Salary: Full time	$19,500	Request from Embassy
Contractual	Website: build and host	$2,000	Request from Embassy
Supplies	Data acquisition	$1,000	Request from Embassy
Supplies	Research	$1,000	Request from Embassy
Other	Translation	$1,000	Request from Embassy
	Total	**$24,500**	**Request from Embassy**

XII. Future Funding or Sustainability

The project is a foundation for future development in the geospatial and temporal distribution of news story contents. The proof of concept should demonstrate the value of such filtering and may be built upon in one or more of the following ways: 1) as a free tool; 2) as the base of a new online news journal product; 3) incorporated into existing online databases (such as those hosted through the Information and Resource Center of the U.S. Embassy's Public Affairs Section) to incorporate the temporal spatial dimension into and enhance their own thematic tools; or 4) to be incorporated into municipalities as a public participation platform / community empowerment tool to better understand incidents that are spatially relevant.

This last option is especially interesting if future planned events and city data are layered in. It is also the direction of most interest to the developer and future funds would likely support an iteration of collaboration with one or more cities to design a public participation tool. Additional functionality may include additional languages, additional study areas, development of a smart phone application, an option for automatic localization (such as for geo-tagging news stories or proximal searching), incorporation of historical datasets, incorporation of future events, additional data visualization options, APIs for integration with other applications, white-labeling options for commercial applications, etc.

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

XIII. Appendix

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 .

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').

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.

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

Gazetteer: A geographical index relating descriptors to location; ex: https://www.geonames.org/ , which related names of places to geographical coordinates.

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

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).

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.

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

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

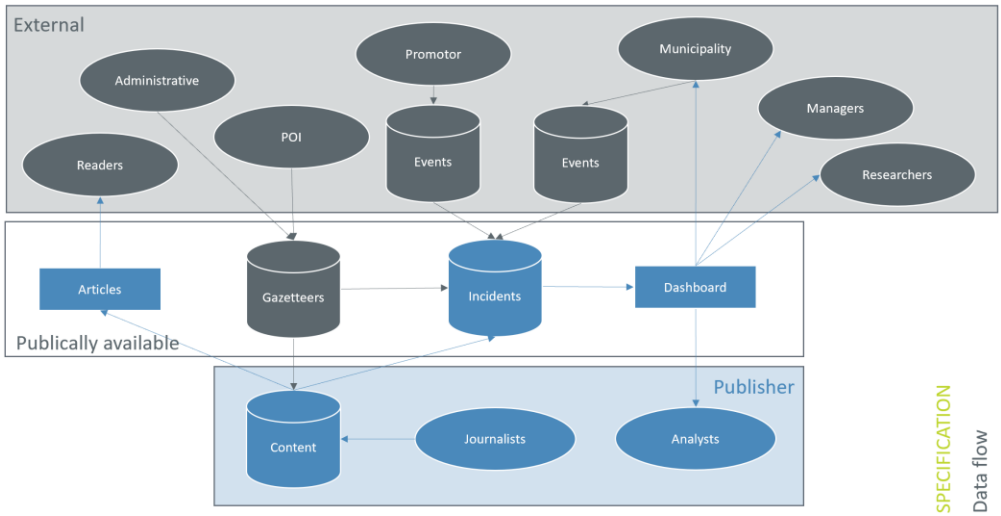
Tag: content, section, or descriptive designations defined by the media publisher; ex: 'política', 'primeiro-ministro', 'governo' (from Público), or 'coronavirus', 'denver', 'homelessness' (from The Denver Post).

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

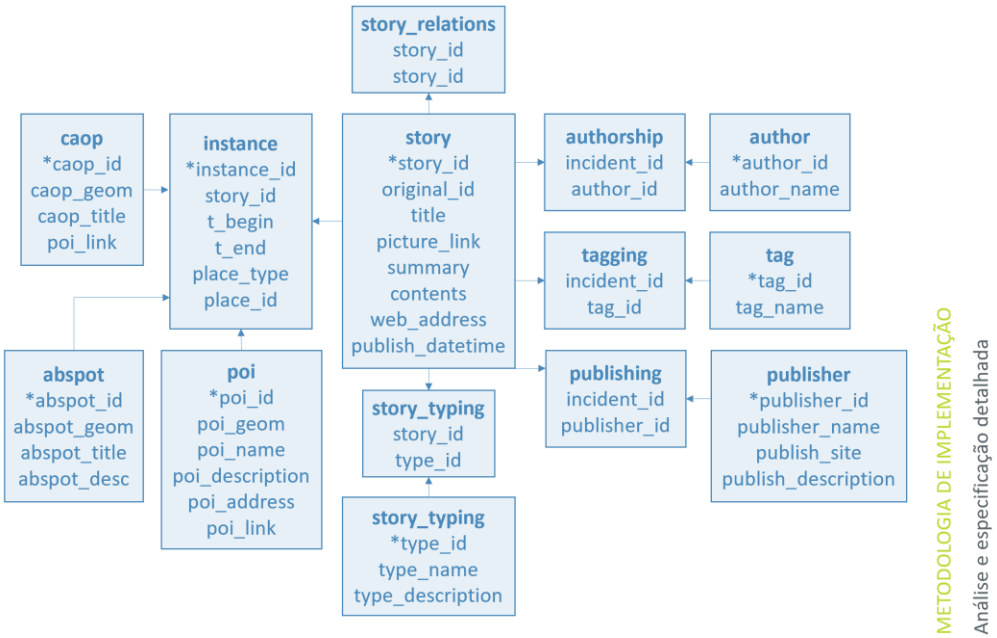
Web application (Web App): a program running on a web server that is accessible via a web browser with internet connectivity.

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

Appendix B: Preliminary data and information flow



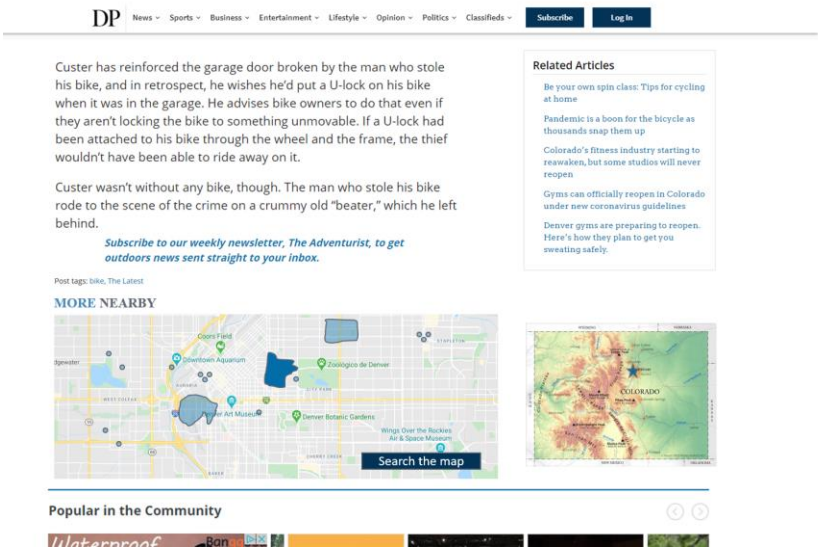
Appendix C: Preliminary data model of the spatial database



Appendix D: Preliminary *Input* layout

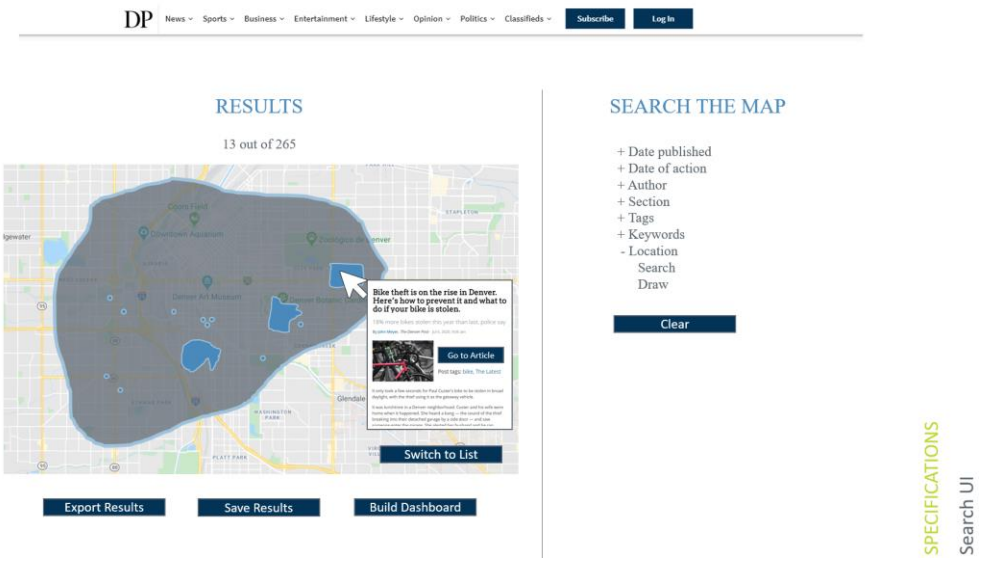
SPECIFICATION Input UI

Appendix E: Preliminary *Context* layout

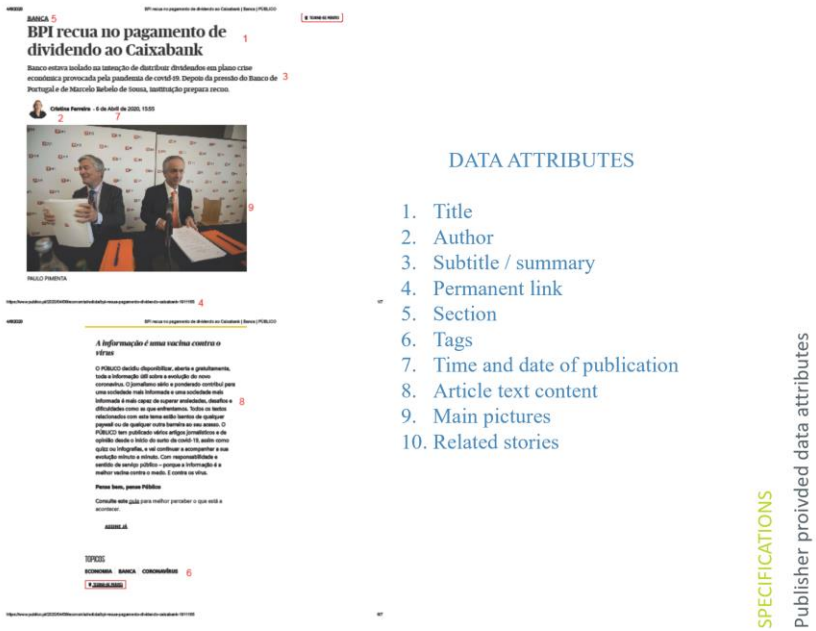


SPECIFICATIONS Context Layout

Appendix F: Preliminary *Search* layout



Appendix G: Publisher provided data attributes



Appendix H: Existing efforts

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 others) 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 subsection of two or more). 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 half mile buffer) and search for all spatially related results, nor easily 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.

Appendix I: Relevant coursework

Cartographic sciences, geographic information standards, geospatial intelligence (GEOINT), geo-statistics, geospatial data mining, modeling in GIS, GIS in organizations, open software and programming in GIS, geographic databases and geospatial web services, geographic information systems, information technology in cities (I and II), mobile and ubiquitous computing, sustainable cities, urban analytics, remote sensing, cybersecurity, and big data.