An open source spatial news web app development project

Masters in GIS&S Thesis Proposal

Callie Wentling
Advisor: Professor Marco Painho, Ph.D

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 Framework

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. 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.

This project seeks to develop a set of functional tools that supports the creation and management of a spatial database of 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 Section 3 for more details. 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.

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

1.1 Study area

The project will use a study area (news story data from at least one section of a publication for a defined time interval) of Lisbon, Portugal (such as "Local" in Público for Q3 of 2020) In the case that opportunities to include additional study areas such as other cities, 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. By building a tool specific to Lisbon, the project seeks to accommodate the culture and business processes of the local community, providing a platform that is useful and valuable

to users (whether citizens, local officials, researchers, or publishers).

1.2 Languages

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 language options of English and Portuguese should support the international use and cross-investigation of a wider user base.

1.3 Access to results

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.

1.4 Sustainability

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

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

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

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

1.5 Impact

1. 1 webapp, freely and openly accessible, available in English and Portuguese languages
2. 1 webapp development code, open licensed for further or related future development by any individual or organization.

2 Hypothesis

The goal of this project is to explore if there is a value to plotting the spatial distribution of news story contents, as well as provide users a portal through which to interact with the data to extract the desired information.

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. Users will find additional insights from the ability to view or filter spatial attributes, especially in conjunction with thematic and temporal attributes. This type of data preparation, though it is initially cumbersome to establish and requires adjustment of publishers' processes to maintain, will provide a powerful foundation from which future economic (improved publisher products elevating their offering and attracting/maintaining a customer base), societal (illumination of local trends requiring intervention, improved community engagement of readers with their surroundings, or improved city resources), and academic (improved research functionality) benefits may stem. If this type of functionality and improved user experience are well-implemented by a handful of productive news services, it will force a shift of the industry standard towards integration of spatial attributes and spatially related products.

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, Section C) are not included in this endeavor, the resulting tools should provide a basis from which future projects may develop and evaluate.

3 Objectives

The proposed 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 B/Figure 2 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 B/Figure 3 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 B/Figure 4 for preliminarsy *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 B/Figure 5 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.

Beyond the implementation of a POC toolset and demonstration of value, this project also seeks to enhance my skillset and experience in the following ways:

1. Planning and execution of a GIS "product"
2. Creation and maintenance of a geospatial database
3. Programming of user interfaces
4. Leveraging of open source programs and tools
5. Incorporation of multi-lingual functionality
6. Collaboration with news industry users
7. Collaboration with city official users
8. Collaboration with readers users
9. Leveraging the knowledge and experience on smart cities, public participation, and geospatial services at NOVA
10. Development of a "smart city" product
11. Development of open source tools
12. Provisioning of the base product for future expansion into desired directions (integration of future language options, accommodation of multiple news sources, integration of planned events, integration of city resources, integration of automatically extracted place from historic sources, etc.)

4 Methodology

To support the identified objectives, the following must occur:

- Perform literature review of prior art and study of existing relevant platforms and tools.
- Conduct interviews with stakeholders (publishers, journalists, readers, researchers, and city officials) to establish and prioritize functionality elements.
- Finalize specifications, mockups of Web App functionality, data model, and finalization of relevant tools and libraries.
- Initialize the development environment.
- Receive data from collaborating journals within the defined study areas.
- Establish incident database, accommodating multiple language options. Load gazetteer(s) and relevant administrative boundary data.
- Develop and test Input tool.
- Develop and test Search tool.
- Develop and test Dashboard tool.

- Develop and test Context tool.
- Translate Web App content to Portuguese and load translations.
- Migrate site to the server.
- Test among stakeholders.
- Compare against mined location results.
- Compare against existing media search options.
- Document results and plan future development.

5 Preliminary Organization

1. Introduction
   (a) Context
      i. Smart cities [1, 2]
      ii. News
   (b) Hypothesis
   (c) Objectives
2. Literature review
   (a) Web tools for news
      i. News search tools
         A. *Portuguese based publication* (ex: Público) [3]
         B. *US based publication* (ex: The Denver Post) [4]
         C. *Good example*
         D. *Bad example*
      ii. Automatic extraction of place in news
         A. GDELT
      iii. WebGIS as a search tool
         A. AirBnB
         B. UberEats
         C. Idealistia
         D. CML
         E. Facebook events
      iv. Open Source
         A. Definition of terms
         B. Licensing options
      v. Open/free formats in WebSIG
   (b) GIS on the Web
      i. Open/free code solutions
         A. QGIS cloud
         B. ArcGIS Server
         C. Lizmap
         D. Leaflet
         E. MapServer
         F. Geomajas
      ii. Commercial solutions
      iii. Design [5–14]
      iv. Build
         A. Hosting
         B. Front end: Leaflet, openlayers, HTML

C. Backend: Python (shpaely, geopandas, gdal, pyproj), geodjango, geoserver
D. Database: PostGIS, PostgreSQL
E. Data prep: QGIS, ArcMap

(c) Fundamental concepts
i. Map algebra
ii. Map fit
iii. Geotagging data [15, 16]
iv. Application of gazetteers [17–20]
v. Internalization and localization

(d) Summary

3. Study area
(a) Portugal
i. Demographics
ii. Habitual news sources
A. Facebook
B. Newspapers
C. Television
iii. public participation platforms
A. OLX
B. Na Minha Rua

4. Methodology
(a) Data
i. Sources
ii. Categorization of data
iii. Preprocessing
(b) Methods
i. Data model
ii. WebGIS architecture
iii. WebGIS application

5. Results
(a) Resulting tools
(b) Spatial distribution of news in study area
(c) Comparison to traditional methods/automatic extraction

6. Conclusion
(a) Challenges/shortcomings
(b) Future development
(c) Summary

7. Appendix
(a) Models
(b) Specification
(c) User guide
(d) Codebase

6 Project timeline of milestone deliverables

Year	Month	Day	Presentation	Progress
2020	Oct	20	Proposal	Data collection
	Nov	17	Literature review	
	Dec	15	Data and methodology	
2021	Jan	19	Analysis and results	
	Feb	18	Planning	
	Mar			Data analysis, first draft
	Apr			
	May			
	Jun			
	Jul			Draft revision, submission for feedback
	Aug			
	Sep			Final submission
	Oct			
	Nov			

Table 1: Proposed scheduled of development

References

[1] Tiago H Moreira De Oliveira and Marco Painho. Open Geospatial Data Contribution Towards Sentiment Analysis Within the Human Dimension of Smart Cities. In A Mobasheri, editor, *Open Source Geospatial Science for Urban Sudies.*, pages 75–95. Springer International Publishing, 2021.

[2] Prof Stéphane Roche. UrbComp2012_Paper10.pdf. 2012.

[3] PÚBLICO, 2020.

[4] The Denver Post, 2020.

[5] Fabio Parasecoli and Mateusz Halawa. Rethinking the global table. In May Rosenthal and Catherine Flood, editors, *Food: Bigger than the plate*, pages 80–89. Victoria and Albert Museum, London, 2019.

[6] Eric Fisher. Making the most detailed tweet map ever, dec 2014.

[7] Ben Shneiderman. Data Visualization's Breakthrough Moment in the COVID-19 Crisis, apr 2020.

[8] Elijah Meeks. 2019 Was the Year Data Visualization Hit the Mainstream, dec 2019.

[9] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Technical report, University of Maryland, College Park, jul 1996.

[10] 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(1):512–522, 2019.

[11] Michelle Rial. *Am I overthinking this? Over-answers life's questions in 101 charts.* Chronicle Books, San Francisco, United Sates, aug 2019.

[12] C Giovando, N Vanraes, I Kanellopoulos, and G Saio. The nature-GIS portal: a gateway for accessing geospatial data on protected areas and nature preservation in Europe. Technical report, Warsaw, Poland, jun 2004.

[13] Katy Börner, Andreas Bueckle, and Michael Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *PNAS*, 116(6):1857–1864, feb 2019.

[14] Alberto Cairo. *The truthful art: data, charts, and maps for communication.* New Riders, San Francisco, 2016.

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

[16] 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.

[17] Sabine Witschas. Cross-border mapping-Geodata and Geonames. Technical report, 2004.

[18] NGA: GNS Home.

[19] Alexandria Digital Library Gazetteer — UCSB Library, 2019.

[20] Richard Nordquist. Exonyms and Endonyms, jan 2018.

[21] Kostas Kampourakis. *Television and the genetic imaginary.* 2020.

[22] 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(13):1–4, 2013.

[23] The Armed Conflict Location & Event Data Project, 2020.

[24] Matt Makai. Deployment.

[25] OpenStreetMap.

[26] Cláudia Henriques, Pedro Marques Gomes, and Sílvia Marques Torres. OS MÉDIA NO PORTUGAL CONTEMPORÂNEO: da ditadura à democ. *Imprensa da Universidade de Coimbra*, 35:19(2), 2019.

[27] Jerry Low. How to Host Your Own Website: Step-by-step Tutorial, sep 2020.

[28] Ramiz Sami. Tools I recommend for building Geospatial Web Applications — by Ramiz Sami — The Startup — Medium, oct 2019.

[29] Ulrich Drepper, Jim Meyering, François Pinard, and Bruno Haible. GNU gettext tools, version 0.21. Technical report, Free Software Foundation, Inc., jul 2020.

[30] Albert Acedo, Tiago Oliveira, Mijail Naranjo-Zolotov, and Marco Painho. Place and city: toward a geography of engagement. *Heliyon*, 5(8):e02261, 2019.

[31] Karina Popova. Politicization of non-political events: A Geospatial Analysis of Twitter Content During The 2014 FIFA World Cup. 2015.

[32] Stig Hjarvard. *The mediatization of culture and society.* 2013.

[33] 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(Gistam):282–289, 2018.

[34] Mark Fisher. *Capitalist Realism: Is there no alternative?* zero books, 2019.

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: GeoNames , 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.

B Preliminary specification

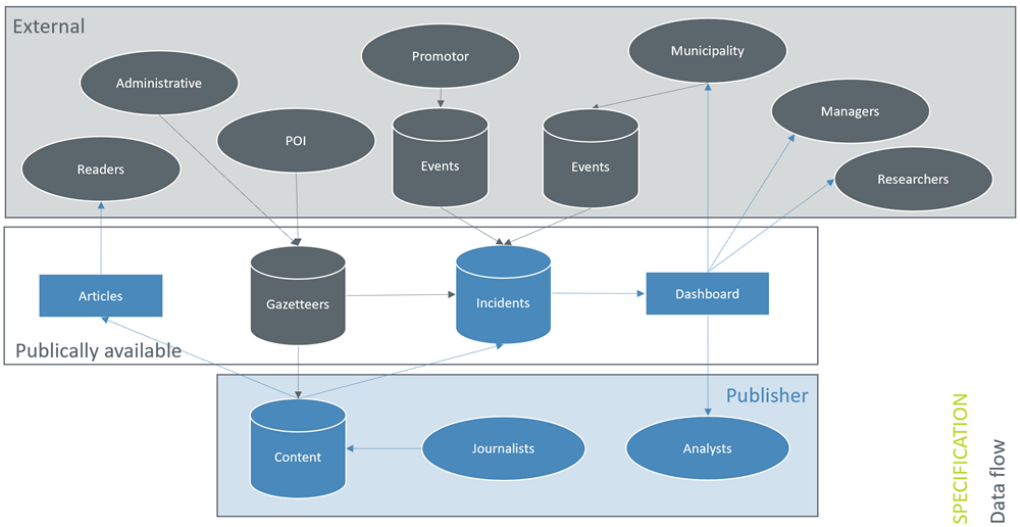

Figure 1: Preliminary data and information flow

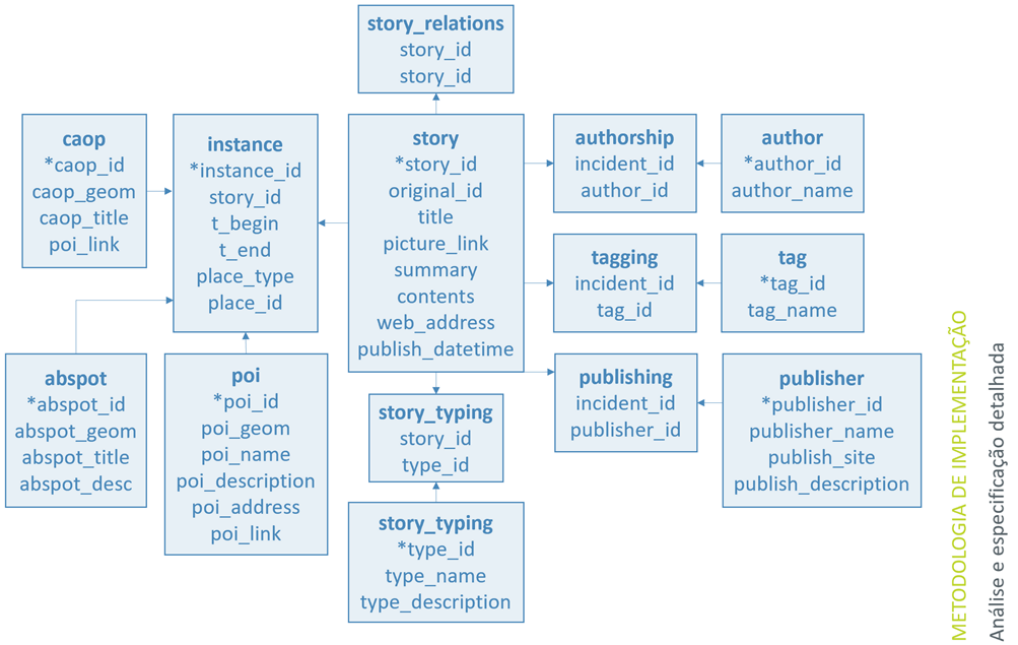

Figure 2: Preliminary data model fo the spatial database

SPECIFICATION Input UI

Figure 3: Preliminary *Input* layout

SPECIFICATIONS Context Layout

Figure 4: Preliminary *Context* layout

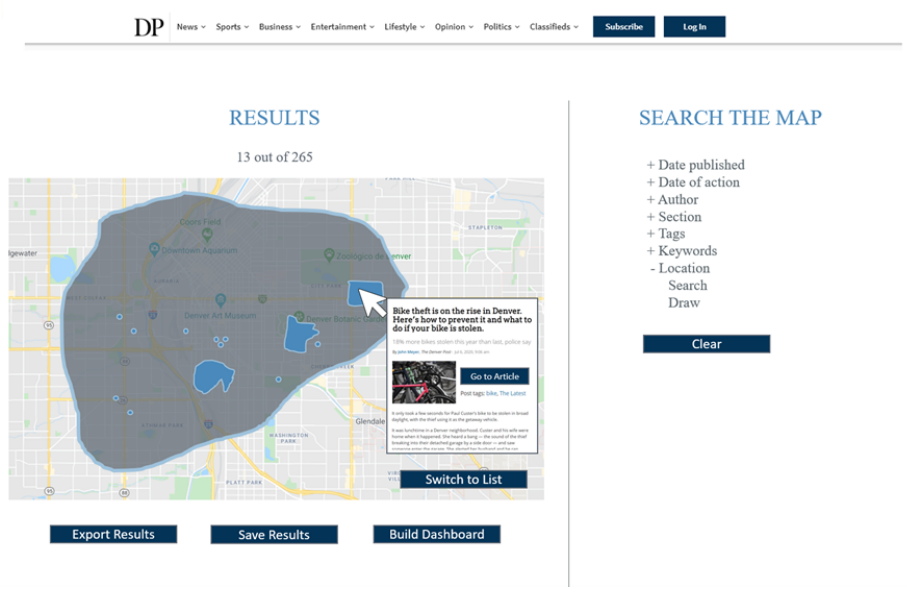

Figure 5: Preliminary *Search* layout

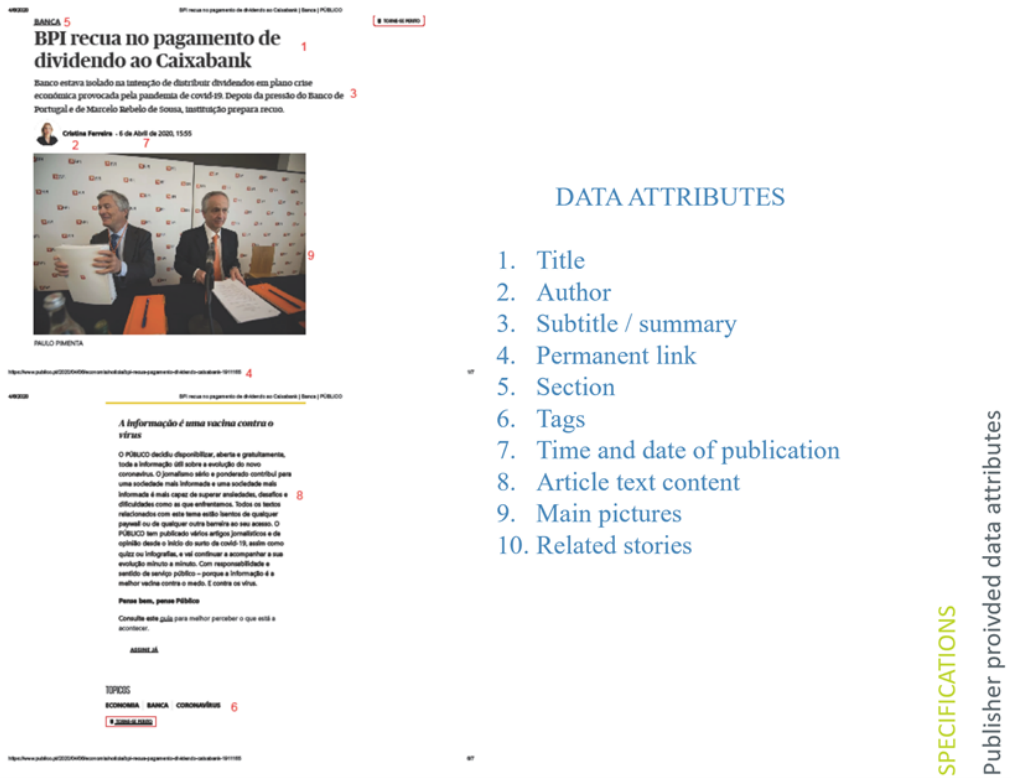

Figure 6: Publisher provided data attributes

C 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 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.

D Relevant coursework

- Cartographic sciences
- Geographic information standards
- Geospatial intelligence (GEOINT)
- Geo-statistics
- Geospatial data mining
- Modeling in GIS
- GIS in organization
- Open software and programming in GIS
- Geographic databases and geospatial web services
- Geographic information system
- Information technology in cities (I and II)
- Mobile and ubiquitous computing
- Sustainable cities
- Urban analytics
- Remote sensing
- Cybersecurity
- Big data