Multi-perspective Approach for Curating and Exploring the History of Climate Change in Latin America within Digital Newspapers

Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Javier A Espinosa-Oviedo, Luis M Vilches-Blázquez

Computer Science and Information Systems 00(0):0000–0000 https://doi.org/10.2298/CSIS123456789X

Multi-perspective Approach for Curating and Exploring the History of Climate Change in Latin America within Digital Newspapers

Genoveva Vargas-Solar[1], José-Luis Zechinelli-Martini[2], Javier A. Espinosa-Oviedo[3], and Luis M. Vilches-Blázquez[4]

[1] CNRS, Univ Lyon, INSA Lyon, UCBL, LIRIS, UMR5205, F-69622 Villeurbanne, France
genoveva.vargas-solar@cnrs.fr
[2] Fundación Universidad de las Américas Puebla, 72820 San Andrés Cholula, Mexico
joseluis.zechinelli@udlap.mx
[3] Univ Lyon, CPE Lyon, LIRIS, UMR5205, F-69616 Villeurbanne, France
javier.espinosa@cpe.fr
[4] Centro de Investigación en Computación, IPN, 07738 Mexico City, Mexico
lmvilches@cic.ipn.mx

Abstract. This paper introduces a multi-perspective approach to deal with curation and exploration issues in historical newspapers. It has been implemented in the platform LACLICHEV (Latin American Climate Change Evolution platform).
Exploring the history of climate change through digitalized newspapers published around two centuries ago introduces four challenges: (1) curating content for tracking entries describing meteorological events; (2) processing (digging into) colloquial language (and its geographic variations[5]) for extracting meteorological events; (3) analyzing newspapers to discover meteorological patterns possibly associated with climate change; (4) designing tools for exploring the extracted content.
LACLICHEV provides tools for curating, exploring, and analyzing historical newspaper articles, their description and location, and the vocabularies used for referring to meteorological events. This platform makes it possible to understand and identify possible patterns and models that can build an empirical and social view of the history of climate change in the Latin American region.

Keywords: data curation, metadata extraction, data collections exploration, data analytics.

1. Introduction

Ninety-seven per cent of climate scientists agree that climate-warming trends over the past century are very likely due to human activities[6]. Some observation reports and studies reveal that the planet's average surface temperature has risen about 2.0 degrees Fahrenheit (1.1 degrees Celsius) since the late 19th century. The hypothesis is that this change has been mainly driven by increased carbon dioxide and other human-made atmospheric emissions.

[5] In Iberoamerica, Spanish has variations in the different countries, even if all Spanish-speaking people can perfectly understand each other.
[6] https://climate.nasa.gov/scientific-consensus/

Technological advances have allowed understanding of phenomena and complex systems by collecting many different types of information. Data collections are exported under different releases with different sizes and formats (e.g., CSV, text, excel), sometimes with various quality features. Tools helping to understand, consolidate and correlate data collections are crucial. Even if there is an increasing interest in analysing digital data collections for performing historical studies on climatologic events, the history of climate behaviour is still an open issue that has not revealed missing knowledge. Long historical data studies could make it possible to compute more complete models of climatic phenomena and the conditions in which they emerged. However, meteorology is a young science that started around the 19th century. It is supported by more or less recent data, making it challenging to run an analysis that can give more historical pictures of climatic evolution and its implications using observations instead of extrapolations. Those willing to promote changes in the behaviour of society and industry to reduce emissions that have a role in climate change must convince civil society of the importance of the challenges. For this reason, our work addressed the problem of collecting and analyzing the history of meteorological events to explore how they were described, lived and perceived by civil society. In this sense, the digitalization of data collections has an increasingly vital role in collecting vast amounts of *hidden* data. Thus, considering that digital archives become more easily accessible every time and contain explicit and implicit spatio-temporal information, researchers in GIScience [18], are becoming aware of these new data sources [10], [9], [34], [41]. Moreover, digital data collections make it possible to have an analytic vision of the evolution of environmental, administrative, economic and social phenomena. In this context, our work deals with data collections that report the emergence of meteorological events (e.g., temperature changes, avalanches, river flow growth, or volcano eruptions). However, the digitized collections have some implicit issues. They are often riddled with Optical Character Recognition (OCR) errors that hamper the performance of information retrieval systems. Therefore, handling OCR errors is one of the two significant problems for information retrieval from collections of historical documents. On the other hand, these sources' problems are related to historical language changes since digitized texts are written in the language of their origin.

This paper proposes an extended description of the Latin American Climate Change Evolution platform called LACLICHEV [37]. The objective of LACLICHEV is to provide an integrated platform to expose and study meteorological events described in historical newspapers that are possibly related to the history of climate change in Latin America. In this sense, we hypothesize that the history (in Latin America) is contained in newspaper articles in digital collections available in national libraries of four countries, namely Mexico, Colombia, Ecuador, and Uruguay. Considering this starting point, LACLICHEV addresses the following issues (see Figure 1):

i First, newspaper archaeology, by chasing articles about climatological events using specific vocabulary to discover as many articles as possible (see the left side of the Figure 1). The challenge is choosing adequate vocabulary to increase the chances of getting articles about climatologic events.

ii Second, once an article talks about a climatologic event, it is tagged with Geo-Temporal metadata specifying what happened, where and when it happened, its duration and geographical extent (see the centre of Figure 1). The objective is to build a climatologic event history of empirical observations.

iii Finally, on top of this history, the objective is to run analytics questions and visualize results in maps given that the content is highly spatial (see the right side of the Figure 1).

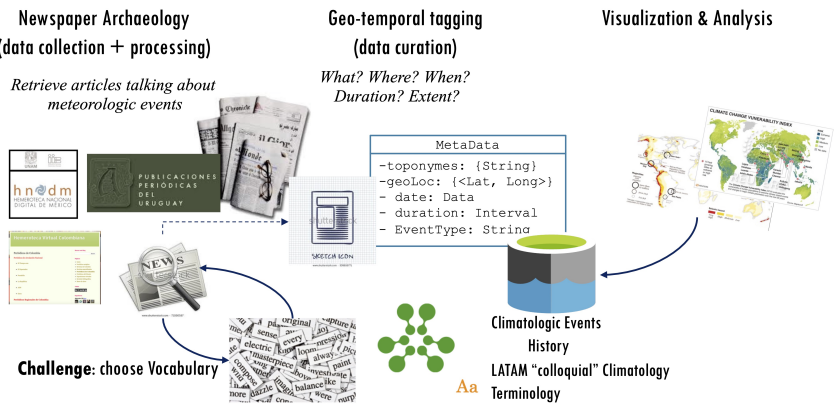

Fig. 1. Problem Statement

Contributions The main contribution of our work is LACLICHEV. It is a data collections exploration platform that applies data collections curation and exploration techniques. These techniques are combined with data retrieval, data analytics, and visualization for understanding the content of articles that report historical meteorological events. These data with high geospatial and temporal content can be aggregated into maps. Maps give a one-shot view of the history of meteorological events observed from the empirical perspective of civil society before the emergence of meteorology as a science [39, 6].
The second contribution is a meteorological event knowledge model that provides several perspectives to describe an event. Perspectives organize metadata that represent such an event is reported through empirical narratives that can appear in newspaper articles, as in the context of our work.
The third contribution is the experimental use of LACLICHEV to build the history of climate change in Latin America from digital newspapers. To track meteorological events, we explored newspapers to search articles that could report such events, the conditions in which they happened, their duration, the places in which they occurred, and their impact in terms of an approximate number of casualties and the kind of damages, etc. As an experimental scenario, we chose the XVIII and XIX centuries, which define a golden age for newspapers in Latin American countries [13], namely, Mexico, Colombia, Ecuador, and Uruguay.

Organisation of the paper The remainder of this paper is organized as follows. Section 2 introduces the general architecture of LACLICHEV and its functions implemented by its

main modules. Section 3 describes the knowledge model we propose for modelling meteorological events as described in empirical narratives written in natural language. Section 4 describes the general curation and exploration processes implemented by LACLICHEV to deal with the curation and exploration of historical newspaper articles potentially reporting on climatologic events. It also describes the use cases that we conducted to evaluate it. Section 5 studies approaches that promote datasets exploration for defining the type of analysis possible on top of them. Finally, 6 concludes the paper underlying the contribution and discusses future work.

2. LACLICHEV for curating and exploring historical newspapers articles

Figure 2 shows the general architecture of LACLICHEV organised into three layers:

i frontend with an interface providing functions for curating articles and creating events descriptions; and giving access to explore the event history containing curated articles reporting meteorological events;
ii backend with the meteorological event history stored in a document management system (see number 1 in Figure 2) and modules for curating (pre-processing and tagging the textual content of newspaper articles - number 2 in Figure 2) and exploring events (see number 3 in Figure 2);
iii external layer connecting to document providers that are available through servers accessible on the Web and APIs exported, for example, by libraries.

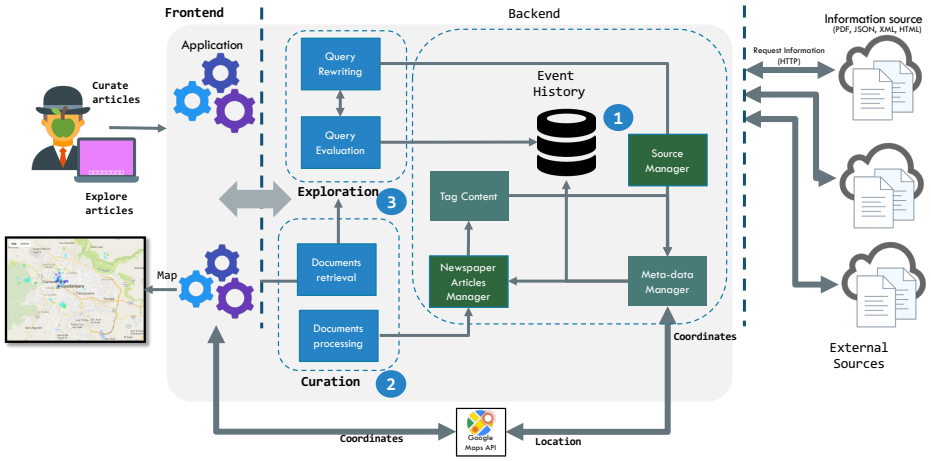

Fig. 2. General functional architecture of LACLICHEV

In the following sections, we describe the core layer of LACLICHEV, namely the backend with its main components, the meteorological event's history, and associated

curation modules used to feed the history and exploration modules to process queries to explore this history.

2.1. Meteorological events history

The event history (storing metadata describing an event from several perspectives, see number 1 in Figure 2) is based on an event knowledge model that we proposed and that is described in Section 3. Through this knowledge model, it is possible to represent the empirical description of a meteorological event with metadata from several perspectives: descriptive (the vocabulary used for describing a meteorological event and the statistics of its use); linguistic (the structure of the sentences used in a narrative describing a meteorological event); the meteorological perspective (represents factual data about an event, location, duration, type, intensity, etc.); and the domain knowledge perspective (meta-data about empirical and factual observations provided by meteorology experts, e.g., the fact that strong rainfall can correspond to more the 75 mm/hr rain).

Metadata is stored in persistence support, a key-value or a document store, depending on the technology adopted by each library. In contrast, the raw documents remain archived in a different server or the same store. LACLICHEV uses a document store (i.e., MongoDB[7]) for storing geo-temporally tagged meteorological events. These events' history provides an interface for performing querying and analytics tasks on top of it. The digital collection can be initially queried by filtering the documents by region, country, or year. Digital libraries offer front-ends for performing this classic information retrieval process. For example, select newspapers published in Uruguay (i.e., geographic filter) between 1800-1810 (i.e., temporal filter). It can also perform analytics queries. For example, locate events during the XIX century, enumerate and locate the most famous meteorological events in the region, and create a heat map of the events in Latin America that happened in the last ten years of the XIX century.

2.2. Curating newspapers content modules

The backend of LACLICHEV includes of a set of modules devoted to implement different operations of data curation (see number 2 in Figure 2). The objective of curating (historical) newspaper articles is to build a meteorological events history that newspaper articles reporting events with metadata, providing as much information as possible about the reported event.

Figure 3 shows the newspapers curation process that is a semi-automatic process devoted to:

- find articles reporting this type of events within digital collections available in existing digital libraries repositories;
- geo-tag interactively and store those articles that actually report such events for building a meteorological event database.

LACLICHEV relies on a knowledge graph that integrates a thesaurus classifying meteorological event types, Wordnet and a glossary defining meteorologic characteristics of meteorological events.

[7] This is a recurrent storage strategy when building databases as a result of processing textual content [32].

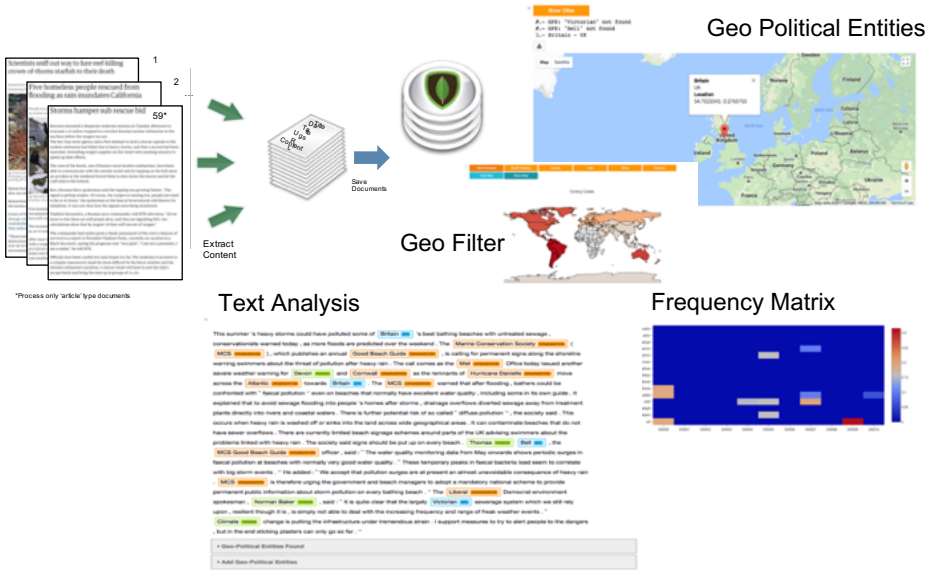

Fig. 3. Newspapers curation process

Curation process Curation tasks can be recurrent and include a human-in-the-loop strategy for validating and adjusting results. For example, suppose an event is geo-tagged to associate it with a geographic location, and the event is described in an article about Montevideo news from Uruguayan newspaper collections. In that case, a human will verify that the geographic location refers to Montevideo in Uruguay and not Minnesota (United States).

During this phase, articles referring to meteorological events are geo-temporally tagged to associate them with the region and/or time window in which they happened. The data analyst validates tags. Since the result can contain a significant number of articles, the user can use three tools to understand the content of the result. The tools let her/him manipulate a terms frequency matrix and heat map.

She/he can also explore the content of the article text using a view that provides information about the context in which the terms are potentially describing an event that appears in the document. For example, the name of geographic locations in the document might refer to the event's location and the region it touched, and a list of geopolitical entities (e.g., school, public building, etc.) to determine the damages caused by the event.

Curation functions provided by the backend modules The data analyst can perform the following curation actions:

- Correct the terms associated with meteorological events that might not be used in such a sense in the text. Indeed, some social and political demonstrations are often described as meteorological events. For a classic automatic text analysis process, this cannot be easy to identify and filter. For example, an article entitled *"Stormy weather within the ails of the*

senate in Ecuador" has nothing to do with the types of events considered but a political one.

- Determine whether personal names correspond to the event's name (e.g. hurricane or storm's name). If that is the case, this information will be used to insert the event into the history.
- Verify whether the names of cities, regions, and countries correspond to geographic entities. The system underlines the names of patronyms, and the data analyst can see the location of the possible geographic entities. Thus, the user can also confirm whether the article refers to the geographic place that she/he is searching for. For instance, if "Santa Clara" is underlined, it can refer to a point of interest, city, or village.
- Determine the date of the event and its characteristics. The temporal terms and adjectives are also underlined to let the data analyst click on those that describe the event.
- Determine the type of damages caused by the event by exploring those terms that describe such information.

The previous actions are used to complete the representation of the articles' content (extracted dynamically) and identify meteorological events more accurately since the data analyst, or domain expert knowledge is used (see Figure 4 showing LACLICHEV interfaces for curation). Note that one event can be described by several articles. In that case, the information stemming from the different sources is loosely integrated by performing the union of the content by applying some rules. For example, suppose the dates reported in two articles do not entirely correspond (variation of the day or the hour). In that case, the date of the event is modelled as an interval computed by processing the dates. If the dates are too disparate, the system keeps a set of dates. A similar process is done with locations; in this case, the system defines a region. A user can define a threshold of the size of a region associated with an event according to its type. Otherwise, the system keeps a set of geographical points.

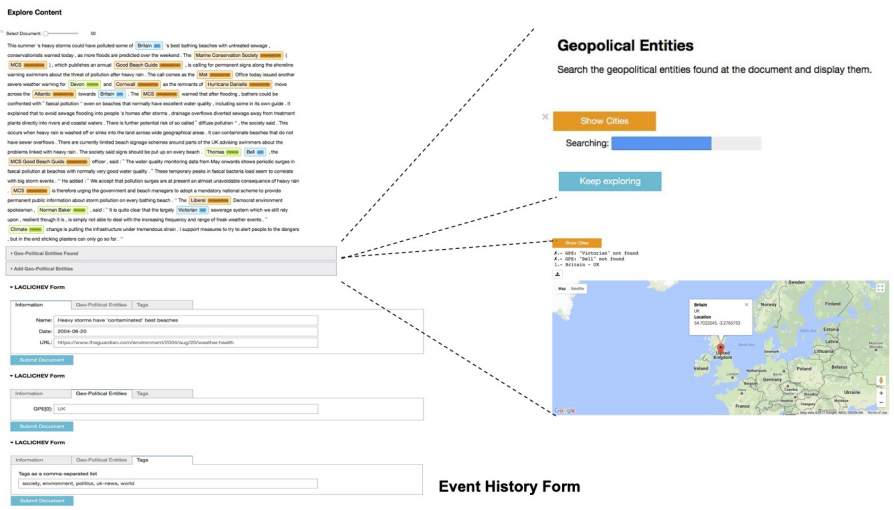

Fig. 4. Event curation process interface for tagging events

2.3. Exploring the collections of digital newspapers

Newspaper articles are explored by conjunctive or disjunctive keyword queries, where keywords can belong to several vocabularies (see number 3 in Figure 2). For example, search articles reporting heavy storms and rivers flooding. The query expressed by a data analyst is automatically completed by using rewriting techniques that consider synonyms, more specific or more general concepts [11]. Thus, three tools can be used for exploring meteorological events depending on expert knowledge of what she/he is looking for.

The rewriting process produces several proposals that the data analyst can adjust and then choose to be evaluated (see details in Section 4). Each chosen query is evaluated using information retrieval techniques, including the article's text stemming for extracting the terms and constructing a frequency matrix that provides occurrence statistics of the representative terms of the text content within a collection of documents.

In general, information retrieval processes do not exhibit this matrix; it is an internal data structure representing the content of the documents and is used to answer queries. In our approach, this frequency matrix is accessible to the data scientist because it provides an aggregated view of the content of a document collection. Additionally, we compute and exhibit a terms heatmap for a given documents collection to provide a more economical (i.e., consolidated) view of the collection's content. Our approach provides an interactive interface that lets data scientists manipulate these data structures to define the piece of collections she/he want to explore.

The data scientist can explore them and then decide whether the collection can describe meteorological events and the documents that might be closer to her requirements. She/he can decide eventually to explore some documents directly or reformulate the query. Once a result containing articles that potentially answer the query has been computed, the user can explore the result and validate the selection elements during the next step of the data exploration workflow.

- *Filtering*. Retrieving factual information, for example, filtering events by region, country or year. For example, Uruguay for the country and between 1800–1810 for the temporal filter.

- *Term frequency*. Understanding the content of digital newspaper collection through the vocabulary used in its articles. Therefore, LACLICHEV exposes the terms frequency matrix and a terms heatmap under an interactive interface. The domain expert can see which are, statistically, the terms most used in the articles, group documents according to the terms used, and choose articles using a specific term.

- *Additional information*. Exploring the content of a specific article using a view that provides information about the name of geographic locations in the document. These locations might refer to the event's location and the region it touched and a list of geospatial features (e.g., school, public building, etc.) to determine, for example, the damages caused by the event.

Exploration process Given a document's collection and associated data structures describing the content of its articles, the data scientist can explore articles to determine whether they report meteorological events. This phase integrates the human-in-the-loop. The reason is that newspaper articles use colloquial terms that can be tricky and refer to metaphors that might not denote a meteorological event. Language subtilities are not easy

to handle manually, mainly because we are dealing with a language used some centuries ago, which increases the challenge of classifying the content of the articles.

3. Meteorological events knowledge model

We propose a meteorological event knowledge model (see Figure 5) to represent climate event reports in digital documents. The objective is to describe events from different perspectives using the information from the articles and newspapers that report them in an empirical form and complete their description with domain knowledge also described in the model. Newspapers do not describe events scientifically; however, we need to locate and profile them by approximating quantitative characteristics to picture the past climate situation in the region. The different perspectives give context to the quantitative features derived/deduced from the descriptions. As shown in Figure 5, events are associated with the newspaper article(s) that describe them (reading from right to left). Each article can have metadata that curates it, pointing to its "raw" content that has been processed and annotated with linguistic labels.

Classes of documents associated with an event (class *Event* in the figure) contain variables that describe its characteristics, like the date it happened or the geographical scope.

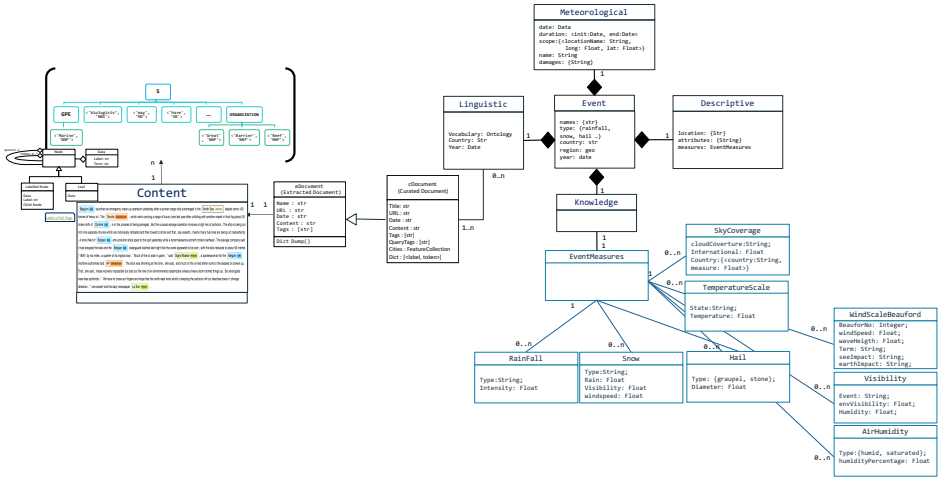

Fig. 5. Event Data Model

According to the perspectives, the event knowledge model provides concepts for representing a *meteorological event*. Each aspect of the knowledge model is implemented using different data structures with associated operations to support exploration actions. The following lines describe the different perspectives of an event and are represented by the event model: descriptive, meteorologic, linguistic and knowledge domain profiles. These perspectives are described in the following sections.

3.1. Descriptive profile

In newspaper articles, there is no generic list of attributes used for describing a meteorological event. Indeed, meteorological events are described in different ways in historical newspaper articles, depending on the author. However, we can often collect information related to location, date, duration, scope, and damages. Meteorological features (like millimetres of precipitations, wind speed, temperature, pressure, etc.) can be explicitly described in articles or deduced according to the description of the event. For example, an event reported in Montevideo describing an overflow of the river implies winds higher than 100 km/h and rain of more than 10 ml/hour, according to the knowledge provided by meteorologists. This knowledge domain is used to complete the reported event's meteorologic features.

We combine scientific knowledge produced a posteriori with empirical observations reported in colloquial narratives centuries before. This strategy can help to estimate the location of the events. Of course, we could have tried a more appropriate approach correlating the location of the event referred to in the article with ancient distributions and organisation of the territory to have a more precise location of the events. For example, we could have looked for the urban distribution of the city in the publication year of the article. Then, compare this result with contemporary maps and have a more accurate location of the events according to the modern urban distribution of the city. For instance, an event reported in Montevideo city's "Rambla" sector in 1910 corresponds to a new quarter today. We will develop this approach in our future work.

3.2. Linguistic perspective

The linguistic perspective gathers the terms used for describing an event in one or several articles belonging to a given newspaper. We propose a tree-based data structure, named *content tree* for representing the content of a historical newspaper article. The tree corresponds to each sentence's grammatical analysis in the article's textual content commonly used in Natural Language Processing (NLP) techniques [4]. The **content tree**, as shown below, consists of a set of sentences. A **sentence** is defined as a set of nodes representing grammatical elements of a sentence and leaves representing the terms composing a sentence in a specific article. We use existing classic NLP techniques because we do not aim at contributing to extending or providing novel ways of using them. The objective is to choose adapted methods for processing the meteorologic newspaper texts.

In Spanish, we use a simplified grammatical model defined by the following simplified Backus-Naur Form (BNF) specification [8]. The simplified specification allowed to process the type of articles we explored, of course an extension of the representation in the next versions of LACLICHEV will allow process other texts describing meteorological phenomena for example in historical novels with narratives about major events:

```
<sentence> ::= <noun-sentence> | <verb-sentence>
<noun-sentence> ::= <named-entity> <conjunction>
                    <noun-sentence>
<noun-sentence> ::= <noun>
```

[8] We have also used a BNF for English to explore the use of LACLICHEV with other languages. This work is out of the scope of this paper and concerns the next version of LACLICHEV.

```
<verb-sentence> ::= <subject> <predicate>
<subject> ::= <article> <noun>
<predicate> ::= <verb> <direct-object>
<direct-object> ::= <article> <noun>
<article> ::= EL | LA | UN | UNA
<noun> ::= "Spanish nouns"
<verb> ::= "Spanish verbs"
```

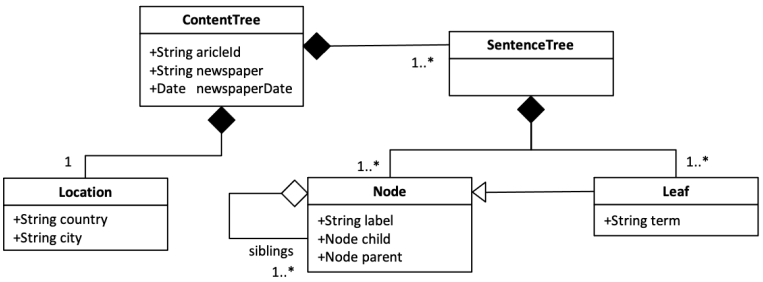

Fig. 6. UML class diagram representing the general structure of a content tree

As shown in Figure 6, a **node** represents a type of grammatical element given in a specific linguistic model defined for a specific language. It is labelled adopting the entity labels produced by classic natural language processing tools known as Part Of Speech (POS) tags. For instance, *noun, proper singular* (NNP), *noun, plural* (NNS), *verb, modal auxiliary* (MD), *Geopolitical entity* (GPE), or Organization. In the case of subjects (NNP), they can be grouped into more general entities that identify geographic locations (GPE), places, names, and organization[9].

A **node** has children, where each child can also be a Node or a Leaf, and a set of siblings, which are other nodes. A **leaf** specializes in a node, and it represents a term contained in the article. A term is a string with a parent, a **node** means a POS tag.

According to the model, the **ContentTree** represents a document's content where the vocabulary used is determined by a **Location** in a country and a city. These classes represent that the same language, Spanish, varies among countries and cities. Recall that in different locations, people describe meteorological events using different vocabulary.

Every article in a newspaper is associated with its content tree. A data analyst or expert domain can explore the articles by navigating their content trees without reading the full content. For example, *retrieve articles reporting heavy storms in Uruguay in December 1810*. Nodes are related through two relation types: instance, correlation. The relation of type correlation describes two terms that appear in the same sentence with a given distance given by the number of intermediate terms.

[9] A full list of POS tags can be found in https://www.cms.gov/

3.3. Meteorological perspective

The meteorological perspective characterizes the event with attributes used to describe it in one or several newspaper articles. Nevertheless, not all the attributes can have an associated value since there might be no evidence within the articles that report it. Attributes, like the date of the event, its geographical scope, or the location of the damaged regions, are computed by navigating through the *content tree* of every article reporting the event.

```
MeteorologicEvent: <date: Data,
                    duration: <init:Date, end:Date>
                    scope:{<locationName: String,
                    long: Float, lat: Float>}
                    name: String, damages: {String}>
```

3.4. Knowledge Domain Perspective

The knowledge domain perspective describes meteorological events using knowledge domain statements created by experts of the National Library of Uruguay. This knowledge has been associated with events through manual analysis of newspaper collections and meteorologists interacting. This knowledge can help interpret the empirical information reported in the articles and complete the information associated with the event description. For example, if the river was flooding due to a storm, it is possible to estimate the wind speed and the approximate litres of rain. The knowledge domain perspective is modelled as a glossary. Figure 7 shows the intuition of its structure.

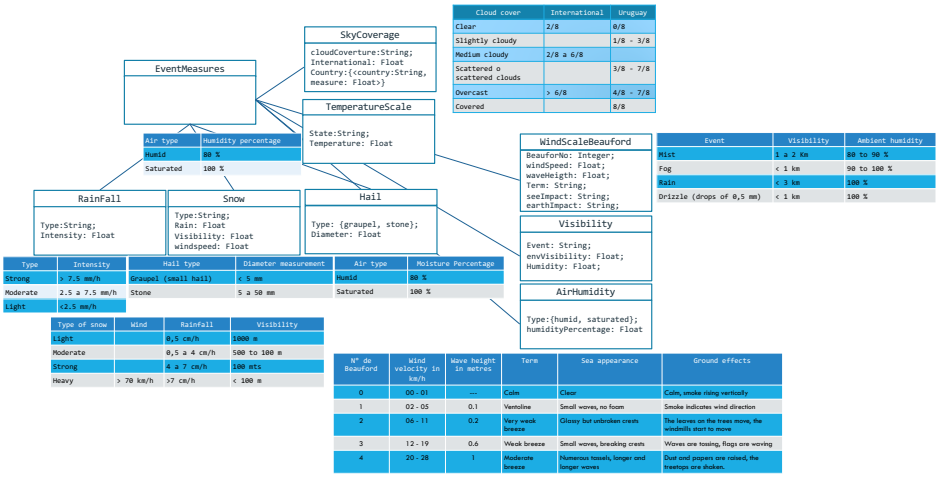

Fig. 7. Climate events glossary

Modelling the empirical knowledge about meteorological events is critical when curating newspapers' descriptions. It represents the interpretation of the emerging content

by observing the phenomena and associating it with metering techniques available today. The principle can be stated as follows: *"today, based on the metrology performed during meteorological events, we know that when the river floods, there is an approximate wind speed and more than "x" litres/meter^2 of rain. So we can estimate the conditions in which the events could have happened in the past.*

4. LACLICHEV in action

LACLICHEV is a client-server system for executing the human-in-the-loop tasks that implement the data exploration process. We have configured LACLICHEV to process historical newspapers of four countries provided by the national libraries of each country. The curated event history has been explored by the librarians of the participating countries. The idea was not to experimentally test the system but to calibrate it according to the characteristics of the digital collections.

4.1. Building a Latin American meteorological event history

We have worked with the national libraries of Mexico, Colombia, Ecuador, and Uruguay to access their newspapers' digital collections. For our experiments, we worked with the collections of the XVIII and XIX centuries of newspapers written in Spanish with the linguistic variations of those mentioned above Latin American countries. The National Libraries of these countries manage historical newspapers with about 4 to 7 million images of newspapers between the XVIII and XIX centuries, depending on the country. For example, the National Library in Mexico maintains 7 million images of digital national newspaper collections. In Colombia the newspaper library is made up of publications published between the end of the 18th century and the first half of the 20th century, including: "El Papel Periódico Ilustrado", "Diario Político de Santafé de Bogotá", "El Alacrán", "El Mosaico", "Semanario del Nuevo Reyno de Granada". It includes newspaper collection from Ecuador and Argentina, namely "La Verdad Desnuda (Guayaquil, Ecuador) and "Vida Intelectual" (Santa Fe, Argentina). The current version of LACLICHEV processed around 19 million images in the newspapers of the fourth countries. The event history has curated 800 different meteorological events.

We curated collections and generated the vocabulary used on articles identified as reporting a meteorological event (see Figure 8). Digital newspaper collections remain in the initial repositories that belong to the libraries. Then, terms and links to the OCR (Optical Character Recognition) archives containing documents with articles reporting meteorological events were stored in distributed histories managed in each country. As shown in Figure 8, the process consists of five steps usually used in natural language processing techniques: sentence segmentation, tokenization, speech tagging, entity and relation detection. LACLICHEV implements these phases in Python, relying on the NLTK library.

The first phase of the pre-processing process of newspapers leads to graphs representing the content of the articles and classic inverse index and frequency matrices used for performing exploration queries.

Besides curating the data collections' content, we wanted to discover linguistic variations in different Latin American countries to describe meteorological events. People's

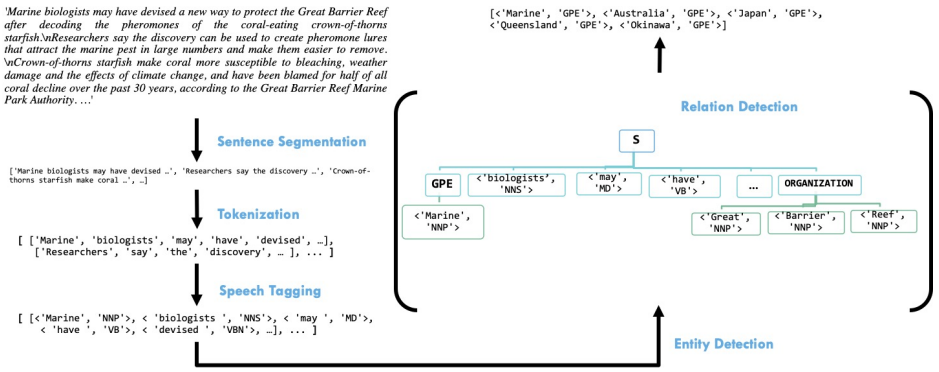

Fig. 8. Pre-processing text pipeline

language and variations can picture civilians' perception of these events, consequences, and associated explanations. Thus, local vocabularies were created out of the terms used in newspapers' articles (see Figure 9). For example, referring to a storm as a stormtrooper[10] Then we updated and enriched through queries, exploration and analytic activities, these vocabularies through human-in-the-loop actions. Data analysts tagged "colloquial" terms used to describe meteorological events and associated them with more scientific terms. These terms can be then used for defining keyword queries for exploring newspaper datasets.

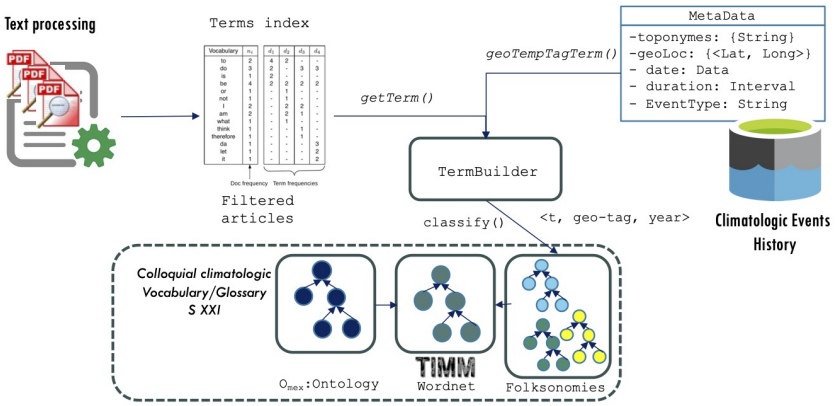

Fig. 9. Collecting colloquial vocabulary

[10] In Mexico, a storm is called a "chaparrón" and in Uruguay, it is called a "chubasco".

4.2. Curating data collections

LACLICHEV proposes functions that data scientists may exploit through diverse functionalities. Next, we present the type of functions of LACLICHEV's API (application programming interface). The implementation of these functions were adapted to the case of historical newspaper collections:

- **Curating data collections** by exploring and processing their content for building the history of meteorological events possibly related to climate change in the considered Latin American countries. The functions for processing texts in Spanish are the core of LACLICHEV. They were coupled with other functions to extract, derive and associate as much data as possible to articles describing meteorological events.

Curation tasks were performed on a collection of textual digital documents with minimum associated metadata, particularly those used by digital libraries that own the collections. Each library adopts its metadata schema, but they generally specify the newspaper's name, the country, the date and number, the number of pages, and the window time in which it circulated. Libraries export the metadata schema used to describe these resources and align them to standards used by digital libraries. For example, the editions of the collection of Uruguayan newspapers were published during the first 10 years of the XIX century.

The curation process generated data structures that provide an abstract representation of the content of each article describing an event. A frequency matrix integrated the terms representing the content of articles extracted from the different libraries' collections. This matrix was sharded and allocated to the servers devoted to interacting with each library. This strategy implies having queries evaluated on different servers. This distributed query evaluation was supported by an inverted index that provided information about the documents containing specific terms and their location. With the inverted index, the curation process also created initial vocabularies, classified by location (country and city) and year. These vocabularies classified the terms used to describe climatologic events in the different Hispanophone countries in LATAM. The temporal dimension allowed to store information about their evolution.

Querying the event's history of already tagged events can be done by keyword oriented queries (e.g., locate the most famous events in Mexico during the XVIII century). Users decide to use some terms that can belong to any of the vocabularies generated in the pre-processing phase. LACLICHEV applies query rewriting techniques to extended user-expressed queries with synonyms, subsuming and general terms. The particular characteristic of this task is that the user (i.e., data analyst) can interact and guide the process according to her/his knowledge and expectations about what she/he expects to explore and search. The first result of this process, based on a "queries as answers approach", is a set of queries that can potentially provide the largest number of results stemming from the collections of the different libraries. The details of the approach we proposed for LACLICHEV is detailed in the following section.

- **Analytics operations and analysis results** are generally presented within maps (e.g., how did rainy periods evolved in the region?). In the current version of LACLICHEV analytics queries cannot be expressed in the frontend. They are implemented manually through notebooks running on top of the event history. The analytics queries concerning

aggregative queries on the event's history, for example, the number of events happening in a country within a specific time window. The average wind speed and millimetres of water per hour deduced for events regarding rainfalls and hurricanes in Montevideo. Classifying the terms used for describing specific types of events. The event history is a curated and clean data collection on top of which other analytics models can be applied for discovering knowledge. This characteristics open analytics perspectives for future uses of LACLICHEV. For example, applying supervised learning for analysing newspapers articles and determining whether they describe meteorological events. This can allow to semi-automatise the curation process and enhance it with a recommendation system.

- **Managing vocabularies**, adding terms, guiding their classification and studying the linguistic connections between the terms used in the different countries. The vocabularies in the current version of LACLICHEV are implemented as RDF ontologies, and it relies on SPARQL mechanisms for querying them. This version mainly addresses the construction of vocabularies and their maintenance as new terms are identified in events' descriptions.

Next subsections describe exploration techniques implemented for meteorological events in the history built through the newspaper articles in the Latin American countries we used.

4.3. Query rewriting

Queries-as-answers exploration. Data analysts can express queries that can potentially explore historical newspaper content to find articles describing meteorological events. The aim is to have a good balance between precision and recall despite the ambiguity of the language (Spanish variations in naming meteorological events). The domain experts must express "clever" queries that can exploit the collections to achieve this goal.

Queries can be initially conjunctive and disjunctive expressions combining terms chosen from the built-in vocabularies or not. Then, queries are rewritten in an expression tree where nodes are conjunction and disjunction operators and leaves are terms, according to an input query expressed as a conjunction and disjunction of terms potentially belonging to a meteorological vocabulary.

Our approach for rewriting queries is based on a "queries-as-answers" process. This technique rewrites user queries into queries that can produce more precise results according to the explored dataset content. Queries as answers proposed by LACLICHEV consist of a list of frequently used queries. Thus, we focus on the following aspects:

Extending query alternatives using hypernyms and synonyms. An initial conjunctive or disjunctive query is rewritten by extending it with general and more specific terms, synonyms, etc. The terms used to express the query are colloquial vocabulary for denoting meteorological events. The rewriting process can be automatic or interactive, in which case the system proposes alternatives, and the user can validate the proposed terms. For example, if the query is *"heavy storms"*, the query can be completed by adding *"heavy stormtrooper"*, *"heavy storm dust"*. It can also be rewritten with synonyms for the adjective *heavy*. In that case, it creates a combinatorial set of rewritten queries.

Note that the colloquial vocabulary stems from the articles of the curated newspapers. As they are curated, the terms used in the articles feed a vocabulary that is first organised in the frequency matrix produced when texts are processed as part of the curation process.

Then we use Wordnet[11] to look for associated terms and synonyms that help address concepts used in different Spanish-speaking countries. We do not translate the query terms to other languages because our digital data collections contain Spanish newspapers. LACLICHEV allows equivalent terms searching to morph a query. For a new term, LACLICHEV generates a node with the operator and then connects the initial term with the equivalent terms in a disjunctive expression subtree. Thereby, more general terms are collected and related to the initial term with these terms in a conjunctive expression subtree. The result is a new expression tree corresponding to an extended query Q_{ExT}. The query morphing algorithm behind LACLICHEV is described in [35].

Extending query alternatives using cultural terms. Use local vocabularies for generating new query expression trees that substitute the terms used in Q'_{ExTi} with equivalent terms used in a target country (e.g., blizzard instead of a heavy storm). This will result in transformed expression trees each one using the terms of a country ($Q"_{ExT1}$ ... $Q"_{ExTj}$) [38].

We call metaphorically "folksonomies" a series of vocabularies created by processing newspaper articles "local" vocabulary. We make and feed each vocabulary according to the country of origin of the processed newspaper article. This lets us extract the vocabulary used during the XVIII and XIX centuries for describing meteorological events in Latin American countries (i.e. Mexico, Colombia, Ecuador, and Uruguay). Using this information, LACLICHEV can answer the following queries: *How have terms used to describe meteorological events changed between XIX-XX c.? Which are standard terms used to describe meteorological events across Latin American countries? Which is the distance between terms used in XIX-XX c.? Which are the most popular terms used in XIX c. for describing meteorological events?*

Defining filters using knowledge domain. We also use domain knowledge for rewriting the queries. We have a knowledge base provided by domain experts that contains some meteorological event rules. For example, rules state that in the presence of a heavy storm: R1. the wind speed is higher that 118 km/h; R2. the rivers can grow and produce big waves; R3. there are rains between 2,5 7,5 mm/h; R4. the range of surface that can be reached by a 100 km wind speed storm is of 1000 km.

Our approach uses this information to generate possible queries that help the domain expert better precise her/his query or define several queries that can represent what she/he is looking for. For example, the previous initial query "Q_1: heavy storm" is rewritten into new additional queries: "Q_{11}: heavy storm *or storm with wind speed* $>$ *100 km*" (using R1). "Q_{12}: storms with 100 km speed that reached Mexico City" (using R2 and knowing the initial point and geographic information). "Q_{13}: storms touching villages 500 km around Mexico city happening in the same period" (R4). Instead of having a long query expression, our approach proposes queries that the domain expert can choose and combine. Note that the system first generates queries, not answers. The answer to a query is a family of possible queries with some associated samples. The user can then choose those queries that she wants to execute.

A climate glossary associates a term referring to a meteorological event with terms of the LODE ontology [12]). LODE is an ontology for historical publishing events as Linked

[11] http://timm.ujaen.es/recursos/spanish-wordnet-3-0/

[12] http://linkedevents.org/ontology/

Data and physical variables describing events. This information generates new queries, which help users discover more details about historical meteorological events.

Using the climate glossary for transforming Q_{ExT} into queries with terms that can serve as filters. There are variables concerning meteorology concepts in the glossary, like wind speed, rain volume/hour, and the water level of seas, rivers, and lakes. Other variables involve geographic aspects, like the location of an event and the scope of land it reaches. Finally, other variables concern damages caused by a climate event with specific physical and geographic characteristics. These different options generate queries combining variables of the same group and different groups. For example, "heavy storms with winds higher than 150 km/h", "heavy storms with rains higher the 10 mm per square metre", and "heavy storms with rivers' overflow". The result is a set of queries Q'_{ExT1} ... Q'_{ExTj}.

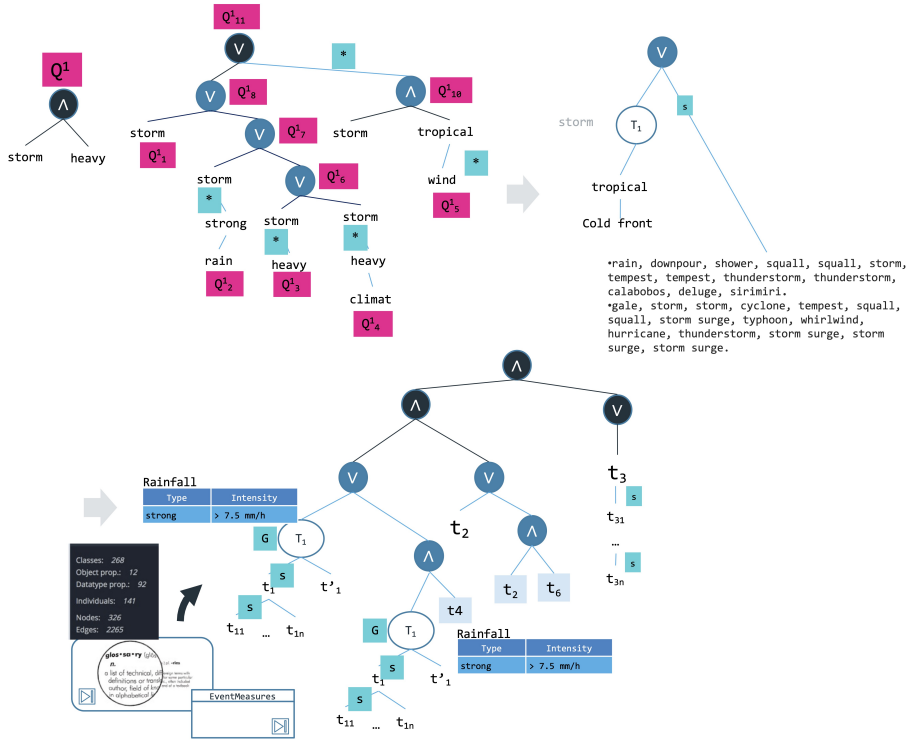

Fig. 10. Queries as answers example

Figure 10 shows an example of the general principle of the queries as answers approach adopted by LACLICHEV. The system rewrites an initial conjunctive query heavy, storm adding concepts (i.e., terms) related to the terms "storm" and "heavy". The figure only shows the rewriting process of the term "storm" for pedagogic reasons. Then in a second round, the system rewrites the query adding synonyms of the terms, as shown in

the upper right side of the figure. Finally, the query is rewritten according to the rules stated in the glossary. LACLICHEV performs a ranking process for the rewritten queries according to the coverage of their potential answer. The queries with the biggest coverage (those that include the largest subset of events in the history database). The algorithms to estimate the coverage of a documents collection are proposed in [35].

4.4. Evaluating queries

The evaluation process of the query is performed first on top of the curated event store. The result is a set of items (events) that answer, to some extent, the query. We also started to generate maps depicting the events reported in the history [6, 39].

Analytics queries LACLICHEV provides and maintains the meteorological event's history on top of which users can visualize information and perform analytical tasks. For example, LACLICHEV can answer spatio-temporal queries like:

- Q_1 *Locate meteorological events in the XVII century*,
- Q_2 *Enumerate and locate the most famous events in the region or in a specific country,* and
- Q_3 *Create a heat map of events in Latin America in the last years of the XIX century*.

The objective is to answer analytic queries that imply aggregating information stored in the event's history. For example, *How did rainy time evolve in time in the region?*, *In which way was climate different between XVII and XIX centuries? How did vocabulary evolve from colloquial to scientific and standardized in the XX century?*

4.5. Scope and limitations

LACLICHEV is running its first version; we expect to enrich the number of digital newspapers digitalised in the libraries. These new items will imply a new curation process that will improve the event history in two directions. First, more articles will describe the already curated events; this will complete the information stored in the history. Second, with more events, we will further test and enhance the analytics queries that require to have a specific volume of data to generate representative maps and analyses about the meteorological events that happened in the past. In future versions, and with more curated events, LACLICHEV is willing to answer prediction queries like *Could it have been possible to predict the evolution of climate behaviour from the data in XVIII and XIX centuries?*. This query requires collecting, curating, and preparing more newspaper articles and other complementary data. However, it concerns future work.
Another limitation of the current LACLICHEV is that it does not provide the adapted mechanisms for exploring the linguistic aspects of the vocabulary. It gathers the terms and organises them in a mesh data structure. Still, it does not provide tools for curating the languages and allowing an analytics exploration of their use of meteorological across countries and time.

5. Related work

Historical analysis of climate behaviour can explain climatologic phenomena and Earth's climate behaviour. There exist several scientific efforts to study the history of climate change. The Climate of the Past [1], for example, is an international scientific journal dedicated to the publication and discussion of research articles, short communications, and review papers on Earth's climate history. The journal covers all temporal scales of climate change and variability, from geological time to multidecade studies of the last century. The Government of Canada provides access to historical observations on climate in Canada starting from 1840 [2]. However, these data collections are disconnected and use different reference variables and observation criteria. They are very heterogeneous and tight to their region. This ad-hoc characteristic is why data curation and exploration processes are essential to extract knowledge that can be digitally analyzed and correlated.

Several domains address aspects that converge in our work, particularly those with certain originality, like data exploration techniques, geographic information retrieval and visualization. The following lines summarise the methods and approaches related to those proposed for LACLICHEV.

5.1. Data curation

Data curation [14, 33] is the art of processing data to maintain it and improve its interest, value, and usefulness through its lifecycle, i.e. improving the quality of the data. Therefore, it implies (i) discovering data collections of interest; (ii) cleaning and transforming new data; (iii) semantically integrating it with other local data collections; and (iv) deduplicating the resulting composites if required. Data curation provides the methodological and technological data management support to address data quality issues, maximizing the usability of the data for analytics and knowledge discovery purposes.

Existing commercial and academic systems provide solutions for curating data [36, 29, 40]. They provide operations for modelling and extracting metadata from raw data collections, and they provide tools for exploring them. Prominent commercial examples are Apache Atlas [13] and SolR [14]. Apache Atlas is a framework for governance and management of metadata. It offers curation functions for metadata typing and classification, data lineage, and exploration functions such as data source search. SolR is a document indexing system including XML files, comma-separated value (CSV) files, data extracted from tables in a database, and files in standard file formats such as Microsoft Word or PDF. Indexing documents can be used as an essential general-purpose curation operation. Its major exploration features include full-text search, hit highlighting and faceted search. Other solutions are built on top of these tools for providing end-to-end general-purpose systems for curating and exploring data, for example, ATLAN [15].

5.2. Data exploration

The emergence of the notion of data exploration provides different perspectives of the data and tools for helping data scientists choose and compound datasets adapted for target

[13] https://atlas.apache.org
[14] https://solr.apache.org
[15] https://atlan.com

experiments [23, 5]. The tools [17] include functions like “data grooming” [27], which denotes transforming raw data into analyzable data with various data structures. Other approaches [24] focus on transforming human-readable data into machine-readable data considering inconsistencies in data formatting given that they are produced under different conditions. The idea is to exhibit processes, digital spaces, and systems that host datasets and provide them with access to understand the conditions in which data are processed.

Data grooming denotes transforming raw data into analysable data with various data structures. Multi-scale queries propose to split a query into multiple queries executed on different database fragments and then perform a union of those queries. This allows scaling the query size as the user gets more confident in her query. Result set post-processing and query morphing go on the premise that the user probably does not need the exact answer to a query. Result set post-processing assumes an array of simple statistical information such as min, max, and mean to be more helpful, especially on massive data sets. Query morphing assumes queries can be wrongly formulated. Query morphing still focuses on answering the query given by the user but will also use a small portion of resources in searching data around the original query.

Query morphing. Another trend regarding data exploration is to tackle the lack of knowledge a user may have on the dataset. Query morphing and queries as an answer are rewriting techniques that compute alternative queries (e.g. adding terms) that can potentially better explore a dataset than an initial query. Approaches such as interactive query expansion (IQE) [30, 8, 19] have shown the importance of data consumers in the data exploration process. Users’ intention helps navigate the unknown data, formulate queries and find the desired information. In most occurrences, user feedback acts as vital relevance criteria for the following query search iteration. The key challenge is identifying bad queries using statistical information or massive scientific databases and identifying interesting queries to return. Identifying bad queries can be done using a list of frequently used queries and returning them based on user feedback.

SVD/PCA [15] is probably the most known algorithm for exploring data sets. It is used to reduce high-dimensional data represented as a matrix. From a practical perspective, it searches for the combination of weighted attributes that expresses the most information, allowing data analysts to work with the more useful 2 or 3-dimensional graphs. From a geometric perspective, these techniques search for the vectors with the highest variance and then express the original matrix according to this new system of dimensions. Using Eigenvalues makes it possible to estimate the amount of information in each dimension [12].

Visualization and summarization are essential to understand the data and maintain it. The field of visual analytics seeks to provide people with better and more effective ways to understand and analyze these large datasets while also enabling them to act upon their findings immediately [22]. Visual analytics provides technology [26, 28] that combines human and electronic data processing strengths. Structured query languages and the graphical interface developed over the top are the standard procedure for accessing data in a database. Many tools exist to perform data visualization with web visualization tools such as D3.js or other tools such as Matlab [20] or R programming language [16]. One of the most critical steps of these tools is to let the data analyst move from confirmatory data

analysis (using charts and other visual representations to present results) to exploratory data analysis (interacting with the data/results). This has led to visual data exploration and visual data mining [7].

5.3. Text processing in newspaper articles

The discovery of knowledge from large-scale text data or semi-structured data is a difficult task that can be addressed with text mining techniques. These techniques extract valuable information to fulfil a user information need. The textual documents available in unstructured and semi-structured forms can be medical, financial, market, scientific, and other documents. Text mining applies a quantitative approach to analyse a massive amount of textual data and tries to solve the information overload problem.

The combination of transformers and self-supervised pretraining has been responsible for a paradigm shift in NLP, information retrieval (IR), and beyond [25]. The approach in [21] extracts target categories, each including many topics. The method extracts word tokens referring to topics related to a specific category. The frequency of word tokens in documents impacts the document's weight calculated using a numerical statistic of term frequency-inverse document frequency (TF-IDF). The proposed approach uses the title, abstract, and keywords of the paper and the categories of topics to perform a classification process. The documents are classified and clustered into the primary categories based on the highest cosine similarity measure between category weight and documents' weights.

The work proposed in [3] discusses the challenge of processing and analysing historical manuscripts. Authors investigate how deep learning models detect and recognise handwritten words in Spanish American notary records. For dealing with natural language (ancient Spanish), professional historians prepared a labelled dataset of 26,482 Spanish words employed in the experiments. The paper [31] proposes a tool that uses raw Spanish text and Spanish event coders for analysing political news articles. The work combines natural language processing techniques, including deep learning and encoders, with the knowledge represented in ontologies to support the automated coding process for Spanish texts.

5.4. Geographic information retrieval

Within GIScience domains, some approaches have developed. [10] and [9] combined methods from Geographic Information Retrieval (GIR) and geovisual analytics to obtain new insights from a digital dictionary about the history of Switzerland. In addition, the authors include sentiment analyses to assess how (historical) places were referred to in texts over time and provide ways to access and explore spatio-temporal information contained in many text archives. [34] described a method to supplement existing records of landslides in Great Britain by searching an electronic archive of regional newspapers. Moreover, the authors construct a Boolean search criterion by experimenting with landslide terminology for four training periods. It allowed the discovery of some spatio-temporal patterns of additional landslides identified in newspaper articles. [41] presented a text-mining program that extracted keywords related to floods' geographic location, date, and damages from newspaper analyses of flash floods in Fujairah, UAE, from 2000 to 2018. Furthermore, this work performed geocoding and validating flood-prone areas generated through Geographic Information System (GIS) modeling.

5.5. Discussion

Any query and analysis must be based on a good understanding of the available data collections because the way they are combined and analyzed impacts the quality and accuracy of the results.

Existing solutions are not delivered in integrated environments that data analysts can comfortably use to explore data collections. The technical effort is still necessary to combine several tools to explore and process datasets and go from raw independent data sets to knowledge, for example, on climate change. Therefore, our research aims to tailor a data exploration environment to help explore digital data collections using a human-in-the-loop approach. In existing solutions, data analysts cannot comfortably explore data collections and design analytics settings, particularly in cases where documents and questions combine scientific observations with empirical observations, like in the case of meteorological events described empirically in the past.

The current version of LACLICHEV did not explore the linguistic aspect, with original or more advanced methods studying texts and combining present and past observations to try to derive conclusions, for example, about climate change.

A technical effort is still necessary to combine several tools to explore and process datasets and go from raw independent data sets to knowledge, understanding and prediction, for example, on climate change. Therefore, LACLICHEV aimed to tailor a data exploration environment that could help explore digital datasets using a human-in-the-loop approach.

Regarding the qualitative assessment of LACLICHEV, we have not run user experience testing to collect feedback and user experience, and we might perform such testing in the future. For the time being, we focus on the analytics such as correlating different descriptions of the "same" event from articles in various newspapers, the location of meteorological events in old maps and their correlation with modern maps. We are working on creating historical cartography of meteorological events that can be confronted with contemporary perceptions of such events.

6. Conclusion and future work

The democratisation of access to data collections opens possibilities for exploring content produced over the years and extracting knowledge that can contribute to understanding critical phenomena like climate change. Rather than directly querying collections for searching documents or performing data analytics operations (statistics, correlations), the objective is to let data scientists understand the content of the collections and then decide what kind of queries to ask. Data exploration is a complex and recurrent process that includes calibrating a querying strategy (defining queries as answers) that can increase the scope of content that can be retrieved and possibly analysed to extract evidence around hypotheses or claims. This new paradigm calls for data curation strategies that are well adapted to describe the content of collections with the right metadata and abstractions.

Our work contributes to data curation and exploration adapted for Spanish textual content within digital newspaper collections. Using well-known information retrieval and analytics techniques, we developed a data exploration environment named LACLICHEV that provides tools for understanding the content of collections. We used digital newspaper collections for applying such techniques for building and analyzing the history of

meteorological events possibly related to climate change in Mexico, Colombia, Ecuador, and Uruguay. The work reported here is the first step toward this ambitious challenge. We continue enriching data collections, developing and testing solutions for generating and sharing step by step this history.

References

1. Climate of the past: An interactive open-access journal of the european geosciences union. http://www.climate-of-the-past.net, european Geosciences Union, Accessed: 2021-04-23
2. Historical climate data. http://climate.weather.gc.ca, government of Canada, Accessed: 2021-04-23
3. Alrasheed, N., Prasanna, S., Rowland, R., Rao, P., Grieco, V., Wasserman, M.: Evaluation of deep learning techniques for content extraction in spanish colonial notary records. In: Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia heritAge Contents. pp. 23–30 (2021)
4. Amavi, J., Ferrari, M.H., Hiot, N.: Natural language querying system through entity enrichment. In: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium. pp. 36–48. Springer (2020)
5. Amer-Yahia, S., Koutrika, G., Bastian, F., Belmpas, T., Braschler, M., Brunner, U., Calvanese, D., Fabricius, M., Gkini, O., Kosten, C., et al.: Inode: building an end-to-end data exploration system in practice [extended vision]. arXiv preprint arXiv:2104.04194 (2021)
6. Ballari, D.: Visualizar datos georreferenciados, http://rpubs.com/daniballari/ipgh-visgeoref
7. Battle, L., Stonebraker, M., Chang, R.: Dynamic reduction of query result sets for interactive visualizaton. In: 2013 IEEE International Conference on Big Data. pp. 1–8 (Oct 2013)
8. Belkin, N.J.: Some (what) grand challenges for information retrieval. In: ACM SIGIR Forum. vol. 42, pp. 47–54. ACM New York, NY, USA (2008)
9. Bruggmann, A., Fabrikant, S.I.: How does giscience support spatio-temporal information search in the humanities? Spatial Cognition & Computation 16(4), 255–271 (2016)
10. Bruggmann, A., Fabrikant, S.I., Janowicz, K., Adams, B., McKenzie, G., Kauppinen, T.: Spatializing a digital text archive about history. In: CEUR Workshop Proceedings. pp. 6–14. No. 1273, CEUR-WS (2014)
11. Carvalho, D.A.S., Souza Neto, P.A., Ghedira-Guegan, C., Bennani, N., Vargas-Solar, G.: Rhone: A quality-based query rewriting algorithm for data integration. In: New Trends in Databases and Information Systems. pp. 80–87. Springer International Publishing, Cham (2016)
12. Chawla, S., Zheng, Y., Hu, J.: Inferring the root cause in road traffic anomalies. In: 2012 IEEE 12th International Conference on Data Mining. pp. 141–150 (Dec 2012)
13. Comesaña, D., Vilches-Blázquez, L.M.: A study of the latin american newspapers from xix-xx centuries with a focus on meteorological events. Revista de historia de América (156), 29–59 (2019)
14. Curry, E., Freitas, A., O'Riáin, S.: The role of community-driven data curation for enterprises. In: Linking enterprise data, pp. 25–47. Springer (2010)
15. Feldman, D., Schmidt, M., Sohler, C.: Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In: Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms. pp. 1434–1453. SIAM (2013)
16. Foundation, T.R.: The r project for statistical computing. (2018), https://www.r-project.org
17. Gewers, F.L., Ferreira, G.R., Arruda, H.F.D., Silva, F.N., Comin, C.H., Amancio, D.R., Costa, L.d.F.: Principal component analysis: A natural approach to data exploration. ACM Computing Surveys (CSUR) 54(4), 1–34 (2021)

18. Goodchild, M.F.: Giscience, geography, form, and process. Annals of the Association of American Geographers 94(4), 709–714 (2004)
19. Goswami, P., Gaussier, E., Amini, M.R.: Exploring the space of information retrieval term scoring functions. Information Processing & Management 53(2), 454–472 (2017)
20. Inc., T.M.: Matlab (2018), https://www.mathworks.com/products/matlab.html
21. Jalal, A.A., Ali, B.H.: Text documents clustering using data mining techniques. International Journal of Electrical & Computer Engineering (2088-8708) 11(1) (2021)
22. Keim, D.A.: Visual exploration of large data sets. Commun. ACM 44(8), 38–44 (Aug 2001), http://doi.acm.org/10.1145/381641.381656
23. Kersten, M.L., Idreos, S., Manegold, S., Liarou, E., et al.: The researcher's guide to the data deluge: Querying a scientific database in just a few seconds. PVLDB Challenges and Visions 3(3) (2011)
24. Kumar, M.S., Rajeshwari, J., Rajasekhar, N.: Exploration on content-based image retrieval methods. In: Pervasive Computing and Social Networking, pp. 51–62. Springer (2022)
25. Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. Synthesis Lectures on Human Language Technologies 14(4), 1–325 (2021)
26. Liu, X., Alharbi, M., Best, J., Chen, J., Diehl, A., Firat, E., Rees, D., Wang, Q., Laramee, R.S.: Visualization resources: A starting point. In: 2021 25th International Conference Information Visualisation (IV). pp. 160–169. IEEE (2021)
27. Liu, Y.: Exploring a corpus-based approach to assessing interpreting quality. In: Testing and Assessment of Interpreting, pp. 159–178. Springer (2021)
28. Mohammed, L.T., AlHabshy, A.A., ElDahshan, K.A.: Big data visualization: A survey. In: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). pp. 1–12. IEEE (2022)
29. Rehm, G., Bourgonje, P., Hegele, S., Kintzel, F., Schneider, J.M., Ostendorff, M., Zaczynska, K., Berger, A., Grill, S., Räuchle, S., et al.: Qurator: innovative technologies for content and data curation. arXiv preprint arXiv:2004.12195 (2020)
30. Ruthven, I.: Re-examining the potential effectiveness of interactive query expansion. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. pp. 213–220 (2003)
31. Salam, S., Khan, L., El-Ghamry, A., Brandt, P., Holmes, J., D'Orazio, V., Osorio, J.: Automatic event coding framework for spanish political news articles. In: 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). pp. 246–253. IEEE (2020)
32. Sodré, A.P., Floriano, L.E.M., Magalhaes, D., Aguiar, C.D., Pozo, A., Hara, C.S.: Comparing alternative storage models for words extracted from legal texts. In: Anais Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados. pp. 36–42. SBC (2021)
33. Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A., Xu, S.: Data curation at scale: the data tamer system. In: Cidr. vol. 2013 (2013)
34. Taylor, F.E., Malamud, B.D., Freeborough, K., Demeritt, D.: Enriching great britain's national landslide database by searching newspaper archives. Geomorphology 249, 52–68 (2015)
35. Vargas-Solar, G., Farokhnejad, M., Espinosa-Oviedo, J.: Towards human-in-the-loop based query rewriting for exploring datasets. In: Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference (2021)
36. Vargas-Solar, G., Kemp, G., Hernández-Gallegos, I., Espinosa-Oviedo, J., Da Silva, C.F., Ghodous, P.: Demonstrating data collections curation and exploration with curare. In: EDBT/ICDT Conference 2019. p. 4 (2019)
37. Vargas-Solar, G., Zechinelli-Martini, J., Espinosa-Oviedo, J.A., Vilches-Blázquez, L.M.: LACLICHEV: exploring the history of climate change in latin america within newspapers digital collections. In: Bellatreche, L., Dumas, M., Karras, P., Matulevicius, R., Awad, A.,

Weidlich, M., Ivanovic, M., Hartig, O. (eds.) New Trends in Database and Information Systems - ADBIS 2021 Short Papers, Doctoral Consortium and Workshops: DOING, SIMPDA, MADEISD, MegaData, CAoNS, Tartu, Estonia, August 24-26, 2021, Proceedings. Communications in Computer and Information Science, vol. 1450, pp. 121–132. Springer (2021), https://doi.org/10.1007/978-3-030-85082-1_11
38. Vargas-Solar, G., Zechinelli-Martini, J.L., Espinosa-Oviedo, J.A.: Computing query sets for better exploring raw data collections. In: 2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). pp. 99–104. IEEE (2018)
39. Vilches-Blázquez, L.M., Ballari, D.: Unveiling the diversity of spatial data infrastructures in latin america: evidence from an exploratory inquiry. Cartography and Geographic Information Science 47(6), 508–523 (2020)
40. Visengeriyeva, L., Abedjan, Z.: Anatomy of metadata for data curation. Journal of Data and Information Quality (JDIQ) 12(3), 1–30 (2020)
41. Yagoub, M., Alsereidi, A.A., Mohamed, E.A., Periyasamy, P., Alameri, R., Aldarmaki, S., Alhashmi, Y.: Newspapers as a validation proxy for gis modeling in fujairah, united arab emirates: identifying flood-prone areas. Natural Hazards 104(1), 111–141 (2020)