Masters Program in Geospatial Technologies

Geospatial database generation from digital newspapers: Use case for risk and disaster domains.

Julio César Preciado López

Dissertation submitted in partial fulfillment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

UNIVERSIDADE NOVA ISEGI
UNIVERSITAT JAUME·I
WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

Geospatial database generation from digital newspapers: Use case for risk and disaster domains.

By Julio Preciado

Supervised by
Rafael Berlanga Llavori PhD
Dept. Lenguajes y Sistemas Informaticos Universitat Jaume I, Castellón, Spain.

Cosupervised by
Carsten Keßler PhD
Institute for Geoinformatics Westfälische Wilhelms-Universität, Münster, Germany.
and
Miguel Neto PhD
Instituto Superior de Estatística e Gestão da Informação Universidade Nova de Lisboa, Lisbon, Portugal.

To my family
Maria Elena,
mecama,
Laura,
lugerarma,
and
GP.

Aknowledgements:

I wish thank Prof. Dr. Werner Kuhn and Dr. Christoph Brox of the Institute for Geoinformatics, University of Münster, Germany; Prof. Dr. Joaquín Huerta and Dr. Michael Gould, of the Dept. Information Systems, University Jaume I, Spain; and, Dr. Marco Painho and Prof. Dr. Fernando Bação, ISEGI, Universidade Nova de Lisboa, Portugal; for granted me with the Erasmus Mundus Scholarship.

My special gratitude goes out to Dr. Rafael Rafael Berlanga Llavori of the Temporal Knowledge Bases Group, University Jaume I , for his steadiness, guidance and most of the technical implementation of this work.

Also I wish thank my co-supervisors Carsten Keßler and Miguel Neto for their comments and suggestions.

I specially acknowledge to Dr. Oralia Oropeza of the Institute of Geography, of the University of Mexico (UNAM), for discussing ideas about the application domain.

Also I would like to thanks Arturo Torres of the National Commission for the Knowledge and Use of Biodiversity and Laura Díaz of the Interactive Visualization Center, University Jaume I, for their comments, suggestions and ideas regarding geospatial services and technologies.

I wish to thank all his my friends at home for being an important part of my life, and my master partners for make the program so pleasant; specially Mateu Aragó.

Acronyms and Abbreviations

BBC British Broadcast Company

DGA - Dependency Grammar Analysis

GeoFeed - Web Feeds tagged with geographical information.

GIE - Geographic Information Extraction

GIR - Geographic Information Retrieval

IE - Information Extraction

IR - Information Retrieval

NER - Named Entity Recognition

NGF - National Geostatistical Framework

NLP - Natural Language Processing

RE - Regular Expressions

RSS - Really Simply Syndication

TDB - Thematic Database

XML - eXtensible Markup Language

Abstract

The generation of geospatial databases is expensive in terms of time and money. Many geospatial users still lack spatial data. Geographic Information Extraction and Retrieval systems can alleviate this problem. This work proposes a method to populate spatial databases automatically from the Web. It applies the approach to the risk and disaster domain taking digital newspapers as a data source. News stories on digital newspapers contain rich thematic information that can be attached to places. The use case of automating spatial database generation is applied to Mexico using placenames. In Mexico, small and medium disasters occur most years. The facts about these are frequently mentioned in newspapers but rarely stored as records in national databases. Therefore, it is difficult to estimate human and material losses of those events.

This work present two ways to extract information from digital news using natural languages techniques for distilling the text, and the national gazetteer codes to achieve placename-attribute disambiguation. Two outputs are presented; a general one that exposes highly relevant news, and another that attaches attributes of interest to placenames. The later achieved a 75% rate of thematic relevance under qualitative analysis.

Key words: Geographic information extraction and retrieval; natural language processing; digital newspapers; risk and disaster domains.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

There are multiple lines of research that use geospatial data. A common problem faced by people study them, is the imperilment to find and acquire thematic information in different geographic scales (i.e., state, municipal, urban, district, etc.). Probably the worst case would be when such information does not exist. This occurs in many countries and Mexico lies among them.

Our scope is to crawl newspapers and extract relevant information for the disaster domain. No matter if the news describe a big, mid or minimum event, we would try to record it in a spatial database. We aim to cover whole the Mexican Republic. And we are focusing on tracking records for State granularity level when county and village levels turns highly ambiguous.

We also intend to make the thematic data base (TDB) as much precise as possible to be readily available to any user. Our scope is to generate the TDB ready to be accessible for any geoclient. Before filtering dynamically daily news, we start using a testing file to review the results and feasibility of further system implementation. Regarding the legal restrictions, once we advanced in developing of the system, we are prone to incorporate the most suitable license for open access to the TDB. In the technical context, we will experiment using Web feeds fro extract geo-information.

The next sub-section of the introduction explain further the reasons that moved the author to carry out this research. After, we state the problem statement and a use case to address, followed by the overall goal and specific objectives to accomplish this research. The last part of section we describe the outline of entire document.

1.1 Motivation

The present research is stimulated by the scarcity of thematic geoinformation related to risks and disasters domains in Mexico. There, the impact of disasters caused by conditions of social vulnerability, environmental degradation and natural hazards was over 10,000 people were killed and 10,390 US$ worth of damage within 1980-1990 [Bitrán 2000]). By that time period the estimated losses were about 700 US$ each year. The above figures are based on information mainly gathered from mid and large size hazardous events.

The number of losses keeps rising. The National Center for Disaster Prevention of Mexico (CENAPRED acronym in Spanish), reported in 2008 1,275 million US$ on materials losses and about 1.5 million victims. These damages were caused by minor events. To highlight, 90% had an hydro-meteorological onset (rainfall, flooding and tropical cyclones) [CENAPRED 2009]. This fact suggest disasters caused by

minor events may be equal or even more expensive than those triggered by large-scale events.

Therefore, the occurrences of smaller scale events are a serious accumulative problem, with the damage mainly built in local areas. Small scale events are poorly recorded in national databases [IADB 2005]. Their identification may help to understand the spatial variability of risk.
Our motivation is to increase logs of different size hazardous events and supplement records onto risk-disaster databases.

The analysis of this type of information can lead to spatial patterns to recognize and reveal risk conditions in different places. This would be an opportunity for communities to prepare better before upcoming natural hazards. It would also encourage stakeholders to make appropriate decisions on time. Better decisions and improved social organization could be translated on reduce lost of lives and goods.

The goal of this work is reinforced by contributions gathered from the Natural Disaster Hotspots Project (NDH). This initiative is led by the World Bank and other research centers [Margaret et al. 2006]. HDN developed an overall risk blueprint analysis dealing with multiple hazards. Some outcomes of their second report presume scarcity on subnational data sets. Most spatial risk databases cover limited periods of time and are too generic. This prevent probability estimation of recurrent dangerous events. As the authors mention:

- *"Geographic areas that are identified as hotspots at the global scale may have a highly variable spatial distribution of risk at finer scales" [...]*
- *"Scale affects data availability and quality, comprehensive, better quality data permit more complete, accurate, and reliable identification of hot spots multi-hazard at finer scales of resolution" [...]*
- *"Scale affects the utility of the results. Better data resolution and a richer set of variables contribute to results that are more local scale relevant for risk management planning "[...]*
- *"The global-and local-scale analyzes are complementary"* [Margaret et al. 2006].

Thus, our work would be a complementary effort for data generation. Focused on thematic geoinformation for the risk study at different spatial scales. We think this information will be suitable to be integrated and combined along other relevant data sets. For instance, base layers such as transportation, water networks, socioeconomic, environmental and health variables may be considered when risk analysis is

performed. Most of those baseline layers are typically created by national mapping agencies however they are not completely devoted to producing thematic data-sets.

1.2 Problem statement and use case

A common problem for spatial data infrastructures (SDI) in developing countries is the lacking of data, its access constraints and its reliability [Masser 2007, Nebert 2004]. Hence, we formulated some research questions to be addressed on the development of this work.

The two main questions are: is it possible to automate the creation of thematic domain data bases from the Web? If so, how to ensure its reliability? In answering these questions we hope the resulting information will be useful in many use cases. One of them is described below.

Risk and disaster management are multidisciplinary by nature. These domains involve researches from social, environmental and earth sciences. And I expect the outputs derived from the current experiment are available for everyone. To name a few of potential users of the database are: students, researchers and public officials of several ministries (i.e water, electricity, planning and civil protection, etc.). In this work all of them are called “the user”.

Supposing the goal of a user is to create a risk index as described in the [IADB 2005] report. These kind of statements need to integrate data from more than one source at different temporal and geographic scales. As a minimum input for analyzing these data and creating new outputs, we should solve the questions of ***where***, ***when***, ***how*** and ***what*** disaster has happened.

To know the place ***where*** a disaster has happened may help us to answer further questions. Such as how far, dense or destructed are the surrounding roads committed to that place. This would improve the efficiency management in the immediately response in case of an emergency. Likewise, it would be useful in the reconstruction and preparedness phases. Similarly, **when** an impact occurs will help to record and log that event. This information can be used to track specific hazards and alert people and authorities to a certain period of time more prone to hazards. For instance, during the rainy season.

How and **what** are parameters that may lead to the categorization of the hazard and risk types. For example, certain phenomena are complex in nature. They can be formed by more than one threat. For instance, a tsunami striking the coast of a populated town may have its origins tens of kilometers away because a strong earthquake triggered it. Another example is to know the places where recurrent

floods and landslides occur and determine if they were triggered by heavy rains left by hurricane rain shields or any other specific factor.

We think most of these questions may be answered properly by providing access to thematic spatial data bases (TDB). TDB containing reliable and constantly updated information may be used to perform many series of analyzes. One problem in their creation is the high time and resource cost. Our alternative to overcome with this problem is the automatic TDB generation using online sources; specifically digital newspapers.

1.3 Context

Here, we briefly describe where our work is related with other data collection initiatives for disaster management.
Most geodata users would agree that almost all spatial databases are thematic. The disaster theme concerned in this work has been studied by many institutions. One of the largest worldwide data set has been created by the Center for Research on the Epidemiology of Disasters (CRED). They have standardized and compiled the disaster Emergency Events Database [1] including records from the early 1900's up to date. They offer free and open access to the database through their website. Another worldwide initiative is the hotspots project, driven by The Center For Hazards and Risk Research at Columbia University (CHRR) [2]. Their project aims to contribute to efforts to reduce disaster losses. Their research program is conducting a global-scale, multi-hazard risk analysis focused on identifying key "hotspots" where the risks of natural disasters are particularly high.

Following the global level, under the umbrella of the International Strategy for Disaster Reduction (ISDR), the Hyogo Framework for Action is committed for achieving disaster resilience for vulnerable communities [3]. They provide links to national platforms that may contain disaster databases.

In the regional framework, a recent initiative is Redhum [4], a virtual tool that provides easy access through the Internet to updated humanitarian information from Latin America, allowing better disaster preparation and response. A similar project, driven by the World Food Program, is the Risk Management and Early Warning for Central America [5].

Concerning the online mapping contribution for disaster reduction, some online interactive maps and geovisualization tool including risk geoinformation are: the Global Disease Alert Map [6] and the "explore our planet" project [7], just to name a few. Most of them only provide means for visualizing such spatial data.

At the regional level, one year ago (November 2008), the Pan-American Institute of History and Geography prepared the agreement for the creation of the Pan American Laboratory for Natural Disaster Monitoring [8]. Among their main goals is to promote the development of spatial databases, in order to support decision making and more efficient early warning systems and improve disaster response.

In Mexico, the National Center for Disaster Prevention and other government agencies in conjunction with some domestic universities work together studying the conditions of risk in the country. Some databases, especially at national level, are available upon request.

This work intends to be a complementary to those works and would be able to be linked with most of the mentioned projects.

1.4 Objectives

The major goal of this work is to create an automatic thematic data base for its potential use in risk and disaster management. The specific goals are to:

- extract information from text files using text mining techniques taking the digital newspapers as a source,
- store such information in a spatial data base and
- make it accessible for any geoclient.

1.5 Thesis outline

Section 2 points to related work from information retrieval and extraction. In the same part we introduce Web feeds followed by a series of examples. In section 3 we explain the development of our methodology. In section 4 we summarize the results to finally draw some conclusions and future improvements to this work (section 5).

2 Related work

This chapter starts with the explanation of common terms referred to through the work. Beyond the context described in the introduction this experiment is closely linked to information retrieval and extraction. Both use techniques of data mining to accomplish knowledge discovery. Hence, we also introduce some of the concepts dealing with them, along other concepts closer to geoinformation retrieval **(GIR)** and extraction **(GIE)**. Afterwards, we provide an overview of various works developed on GIR and GIE. We then describe the data format we are going to use for our experiment, followed by an outline of some projects that have used them.

2.1 Domain concepts

The work presented here focuses on collecting data and information related to risk and disaster management for different periods of time. It is convenient to define the concepts related to those subjects to avoid confusion through the reading.

We consent to use the term **risk** defined as *"the probability of harmful consequences or expected loss of lives, people injured, property, livelihoods, economic activity disrupted or environment damaged resulting from interactions between natural or human-induced hazards and vulnerable conditions*" [UNDP 2004]. Risk is conventionally expressed by the combination of the hazard and vulnerability.

[Blaikie et al. 1994] define **hazard** as *"those natural phenomena that can adversely affect different sites in different time scales, with varying degrees of intensity and severity".* Among the most common threats or hazards are earthquakes, floods, droughts and landslides. The same authors express the term **vulnerability** as *"... a combination of the characteristics of a person or group, expressed in relation to exposure to the threat arising from social and economic status of the individual or community concerned"*. The concept of vulnerability is intimately related to social processes in disaster prone areas and is usually related to the fragility, susceptibility or lack of resilience of the population when faced with different hazards [IADB 2005].

A named "**natural disaster**" is a serious disruption triggered by a natural hazard causing human, material, economic or environmental losses, which exceed the ability of those affected to cope.

The previous four defined terms are used generically throughout the entire document. New ones may appear in later sections explaining issues about their specific context.

2.2 GIR and GIE terminology

Our work is associated mainly with Information retrieval and information extraction. The first concerns to archiving and finding information automatically ([Rijsbergen 1979, Singhal 2001]).

Information extraction (IE) is a type of information retrieval (IR) , whose goal is to automatically extract categorized, contextually and semantically well-defined data of a certain domain, from unstructured documents [40]. Information Extraction distillate pieces of information that are salient to the user's needs [Sundheim 1995, 40].

Text documents are data and may become information if they are relevant for users. Text can be extracted from unstructured and structured documents. Conversely to unstructured text, structured blocks define logic structures and may be easily processed by machines. Thus, unstructured texts are more complex to handle but they are spread immensely on Internet (e.g., social networks).

Text analysis has been carried since around half a century using electronic machines [Singhal 2001]. Information retrieval and extraction use natural language processing (NLP) techniques to mine, in this case text patterns [Christopher et al. 2008] [Rijsbergen 1979]. Natural language processing includes other sub-tasks. One of them is named entity recognition. The goals of named entity recognition (NER**)** are locate and identify entities in text and categorize them. To perform that actions NER requires to be engaged on a series of predefined named categories of a given interest ([40] [Rijsbergen 1979]); for example, a predefined word would be ”Fonden”, which is the Spanish acronym of an aid budget for disaster recovery in Mexico. To categorize block of text, NER tokeinzes words. Tokens are also known as lexemes. Therefore, the lexical processor splits sequences of characters into tokens and re-categorize them according semantic rules [Christopher et al. 2008]. There is a wealth NLP and NER tools, see [43]for example.

NER techniques can be classified into three types:

- Knowledge-based Systems
- Systems based on machine learning models and
- Systems that combine both

Our experiment focuses on Knowledge-based Systems. Such systems use rules patterns and grammars to learned from an armed corpus of text built on heuristics [Sundheim 1995]. Methods to recognize capitalization, perform geographical searching in gazetteers or in extensive name lists also rely on that systems [18].

NER techniques are also been used in other approaches for categorizing word-entities in texts.

Alike NER, other natural languages techniques are used to process text corpora. For instance, "part of the speech" (PoS), which aims for categorize each word according its own attributes and the ones of their neighbors, then reclassifies to assign them to the appropriate category. Other approach is syntactic analysis, this method is devoted to detect relationships between atomic entities through a sentence (i.e., words). Syntactic analysis operates upon grammar rules on sentences of a given language. One if its branches is dependency grammar, whose deals to reveal how several elements across the text are related each other but also how those relationships get focusing on syntactic units [Alfonseca 2007]. Syntactic analysis may be performed superficially (shallow, chunking, etc.) and deeply. The result accuracy of each method depends upon the language used. A brief review of the results performed by the used of different NLP techniques can be accessed on [46].

[Alfonseca 2007] recognizes that extracting textual relations, entity recognition and classification, and ontology population are fundamentally the same problem.

We consider geo information retrieval and extraction (GIR/GIE) when focus for gathering georeferenced information. This relies on the geographic information science (GISc). GIR/GIE community have coined terms frequently used. Some of them are described in the paragraphs below.

Toponyms are used to name places over earth surface. Another commonly referred terms **are geoparsing [10]** and **geocoding [11]**. The former refers to identifying geographic content from unstructured content as documents containing text expressed using natural language. The latter means to turn over geographical coordinates from geospatial entities. For instance, postal codes, streets and addresses. A third associated concept is **geotagging**. This refers the process of attaching explicit space and time identification to various media such as photographs, video and other digital documents. The geotag notion has recently been formalized by [Keßler et al. 2009].

All the above concepts are closely related to each other. For instance, if we have a website and want to know the location of its contents, we may use NLP to *parse* its text paragraphs and annotate geocontent on it. If the same text has entities where street names and ZIP codes can be identified, we may use a geocoding service to obtain their geographic coordinates (this service could be included in the geoparsing algorithm). Also, we may find whether the website emits alerts and check if they have attached geographic coordinates. If so, those files then became *geotagged* and now are GeoFeeds [15]. Also the website may have geotagged photographs ready to be displayed on a map (see for example [13]).

Other terms coined by the GIR community are **Geo** and **Non-Geo ambiguity**. Non-Geo ambiguity occurs when place names are the same as common words [Smith 2002]. This metonymy looms in the case of the word “Freeling”. This word can be used rather to refer a place in South Australia or to name of a computational linguistic library .

Geo ambiguity appears when the same term matches with more than one place [Smith and Crane 2001]. For instance, looking for “Las Vegas” in Geonames’ [14] service, returns a list of populated places in Cuba, Honduras, United States and Venezuela.

Most of GIR and GIE experiments uses **gazetteers** in some way. Gazetteers are geographical dictionaries that list geographic placenames and earth features, such as rivers, lakes and roads, and their location. They were traditionally collected and managed by official mapping agencies. However, recently, "new actors" are working actively in doing the job. Examples of them are private companies, academia and many people from the Web 2.0 community. These dictionaries may contain statistical information attached to the geographic features as well as vernacular and deprecated toponyms.

2.3 GIR and GIE experiments

Early approaches in GIR and GIE can be traced back to the early 80’s [Farrar and Lerud 1982] [Vestavik 2003]. Nevertheless, since the early and mid nineties there have been many advances; especially for indexing geographic features. From that time to date many contributions have been made mixing methods for different purposes. We conducted a brief review of some of them and the main techniques they used.
The GIR/GIE objectives and the procedures vary widely. For example, once terms are disambiguated they are used to improve existing gazetteers or create new ones[Twaroch et al. 2008]. Or also to identify the "geospatial extension" on Web pages in accordance with their thematic context. Similarly, others center their attention on visualizing the geo-content of Web sites. Also, there have been some insights into the use of ontologies ([Jones et al. 2002] [Badia et al. 2007, Borges et al. 2003]) and knowledge representation ([Sallaberry et al. 2007]) for dealing with GIR/GIE.

Several of the ideas mentioned in the above paragraph stem from cognitive and language spatial models ([Egenhofer et al. 1991], [Xu 2007, Egenhofer and Mark 1995]). For example those presented by [Egenhofer et al. 1998], which deal with the use of the language to refer to spatial notions of daily life and their differences when they are expressed to query computer programs, like geographic information systems (GIS).

An early and noticeable contribution on GIR was made by [Woodruff and Plaunt 1994]. They developed what was called **G**eoreferenced **I**nformation **P**rocessing **SY**stem (GIPSY). This system was capable of extracting relevant key words and phrases from documents discussing geographic ideas. The parser was created using lexical constructors to handle spatial relationships in the text, and was also supported to looking up words in a sophisticated thesaurus. That work offered a way to understand the meaning contained in the documents. After parsing and extracting the text, this data was computed through some probabilist functions, ranked and indexed. Finally, polygon overlaying using the weights from the previous ranks was performed. Thus, the goal was to output the most relevant areas mentioned in a given document.
In the area of large-scale gazetteers there have been several contributions. For instance: Perseus Digital Library [18], Alexandria Digital Library [Hill 200], Getty [35] and Geonames [Wick and Becker 2007]. They were developed under different approaches and for different objectives; they are widely used and still studied for the GIR and GIE community.

[Hill 200] carried out pioneering work concerning digital gazetteers for the Alexandria Digital Library project. Specifically regarding georeferencing places with geographic footprints. One of the main assets of the author's work is the process of assigning type categories to places.

As the Internet popularity started growing, the emerging number of websites also grew in parallel. Then, new forms of information retrieval were rooted from multiple sources. For example, in the field of toponym disambiguation, a significant contribution was provided by [Smith and Crane 2001]. Their approach was to ripe a method for toponym-disambiguation and evaluate its performance. The tasks were conducted under Perseus Digital Library project (PDL) [18]. To handle disambiguation, PDL algorithms were constructed from *internal* and *external* evidence. The first uses honorifics, generic geographic labels and linguistic environment. The second is backed on gazetteers, biographical information and other linguistic knowledge sources.

After [Smith and Crane 2001] many other approaches has been developed. For example, [Smith 2002] deployed long stand methods to identify and visualize historical events appearing in unstructured documents. For similar purposes, [Smith and Mann 2003] have made some effort towards toponym disambiguation for unmarked place names in historical corpus of texts .

In the same area [Hu and Ge 2007] have used supervised machine learning for toponym disambiguation. They applied geo name entity recognition (Geo-NER) techniques to a corpus of news collection grounded from a national newspaper. Their objective was to find *"identical"*, *"similar"* or *"part of"* geographical entities through an administrative feature gazetteer. They added more functionality to the system to improve toponym identification and tag-ranking. Afterward, they applied statistical and probabilistic classification models to disambiguate all surplus toponyms. One of their experimental outputs produced an accuracy of around 85% for the national level dictionary.

[Leidner 2007] automate toponym resolution. The authors performed placenamed frequency on text to express foot print extension of refereed places based on the meaning of words.

Other ways to parse the Web for various purposes have been found in the literature. [Twaroch et al. 2008] have experimented using vernacular name extraction from social Web sites to increase the placename entries in traditional gazetteers.

[Leidner et al. 2003] experimented grounding geospatial names focusing on extracting spatial information under the question-answering approach. The authors also created a method to visualize the information retrieved.

[Jones et al. 2008]have employed Web scraping to extract and compute uncertainty between naive and official corpus of geographic terms. The goal of the authors was to enrich and improve the quality of gazetteer's vernacular terminology.

[Amitay et al 2004] have carried out some research for exploiting the Web. They implemented a system capable of geotagging Web content such as photographs. The resulting accuracy was up to 80%. The authors also have performed geographic focus identification on websites. After parsing a set of websites they could detect more than 90% of times the extent of those sites.
Other efforts have been made to geotag Web content such as photographs, Flickr [17] is an example. Likewise, [Ahern et al. 2007] showed how to apply heuristics to the photography domain. What they developed was a method to extract georeferenced photographs and display them in their corresponding spatial context. The authors argue that geovisualization tools help users to understand data trends and features, specifically for this kind of media.
In the same line, [Popescu et al. 2009] have analyzed metadata of photographs to deduct time stay and travel paths of voyagers. The authors used a set of georeferenced photographs to perform their analysis. Their major goal was to analyze time and location tags attached to online photographs and extract information related to trips. They compare their final outputs (trips) versus those added by users on travel planning websites [16].
On the other hand [Manov et al. 2003] have used spatial knowledge base (KB) and ontologies for Web sttrapping. Instead of using geographic gazetteers as a source, the authors implemented structured KB into an ontology. Among the advantages of having that ontology are the way of handling relationship between geographic entities for disambiguation and reasoning. As they state, using ontologies increases system flexibility and enable users to customize it That features and other functionality, such as automatic schematic annotation, indexing and retrieval of unstructured and semi-structured content are found in what they called Knowledge and Information Management (KIM). The authors agreed that performing correctness on spurious terms other systems perform better and therefore its precision value increases.

[Zubizarreta et al 2008] have use geoparsing and geocoding to establish a geographic focus (geofocus) in Web sites. The geoparsing process points to an ontology-based gazetteer for place name disambiguation. Followed by a series of filters tailored to increase the precision of geocoding locations. One such filter rates placenames according to their appear frequency text. Finally, the next agent uses the previous values to determine the geofocus of the crawled websites. The results show high precision with 5% error.

[Turner 2008] claims that extracting passively collected information can provide insightful and powerful data-sets. Other approaches aiming ontology population for geographic data are made by [Badia et al. 2007, Borges et al. 2003] and [Jones et al. 2002]. Similarly, the model used by [Sallaberry et al. 2007] uses semantics as backbone

for extracting geo information from unstructured text; their results are suggestive.

As noted, much work has been done in the area of GIR and GIE. The application of NLP and mix of NER methods has yielded satisfactory results [SLINERC 2002]. In some cases the results have achieved almost hundred percent of accuracy (see [Sundheim 1995]). Other studies did not achieve the desired goals. Several of the techniques cited so far use NER in one of their processing step. [Stokes 2008] have performed measurements over different NER approaches on GIR performance to determine their accuracy. That can be taken as a bench mark when implementing future projects.

The next point give examples of studies carried out by fetching Web feeds for the same purpose.

2.4 Web Feeds and GIR/GIE

In this section we explain some works for GIR using Web feeds. To understand it better and since our experiment uses the same format, we started summarizing the features of Web feeds.

Web feed Web Feed is a data format used to give users frequently updated content. They are a simple way to read and write information on the Web [25]. Web feeds are usually automatically created by a Web content management system (CMS). Those systems are due to manage work flow in a collaborative environment. Specifically, Web CMS are able to simplify the publication of Web content; for instance, by using Web feeds. Those feeds can be accessed in a central way by using a feed reader, which is a Web client that aggregates feeds from different Internet sources[41].

Web feeds are XML documents, then they inherit all XML features such as security, structure, extensibility and namespace. There are two Web feed formats: Really Simply Syndication (RSS) and Atom. Both formats are referred to in this work as "Web feed" or simply as "feed". Their development can be tracked back to the late 90's but Atom is more recent. Likewise, both share more than one version. At the time of this manuscript the Atom comes in its second edition (version 1.0). For RSS, the latest version is 2.0 but version 1.0 is still used in many websites. Below we give further explanation of each format.

2.4.1 Atom

Here we give an introduction about the Atom format. Since this a summary, we will briefly review its structure and highlight some of its common elements.

An Atom document is divided in two parts. The first one is the representation of the feed and includes all other member elements. This item is the root of any Atom document and can be identified with the “feed” element. On the other hand, the “entry” element represents exactly one entry. Inside a feed may be one or more “entry” elements containing the metadata. Figure 1 shows an example of an Atom version 1.0. Required elements at feed level are: <title>, <updated> and <id>. They first identify the feed’s title; the updated element is to determine the time when a feed is modified; id is a permanent universal unique identifier included in feed and entry elements.

Other recommended elements for the root level are: <author> and <link>. The former includes the name and contact address of the author. The link specifies a reference from the feed to a Web resource. This element also must be inside each entry in order to be valid under the World Wide Web Consortium **(W3C)** specification. The document in Figure 1 was simplified for better understanding and may not be w3c valid.

```
<?xml version="1.0" encoding="UTF-8"?>
  <feed xmlns="http://www.w3.org/2005/Atom">
    <id>tag:blogger.com,1999:blog-1072940558039948827</id>
    <updated>2009-07-16T09:24:43.303-07:00 </updated>
    <title type="text"> Atom example</title>

    <link rel="http://schemas.google.com/g/2005#feed"
type="application/atom+xml"
href="http://mapasdf.blogspot.com/feeds/posts/default"/>

    <author>
        <name> Author name </name>
        <uri> http://www.blogger.com/profile/00317610861555096095</uri>
        <email> noreply@blogger.com </email>
    </author>
  <entry>
    <id>tag:blogger.com,1999:blog-1072940558039948827.post-
7021818904540026515</id>
    <published> 2009-07-16T08:02:00.000-07:00</published>
    <updated> 2009-07-16T08:04:41.197-07:00</updated>
    <title type="text">Working with Atom feeds!</title>
    <content type="html"> Her can be stored text, images, audio and video. “Type”
attribute is now set for html content.</content>
    <linkrel="replies"type="application/atom+xml"
href="http://mapasdf.blogspot.com/feeds
/7021818904540026515/comments/
default"/>
  </entry>
  </feed>
```

Figure 1: Sample structure and content of an Atom 1.0 document.

Mandatory Atom elements are emphasized in italics. The entry element (bold and italics) define the metadata of the feed. Every entry has a unique id, publishing and updating elements. The content element is hatched for storing text, audio, images, video or other any other media. The link element may have up to six different attributes. In our example (Figure 1) it is noticeable that the “href” attribute appears several times. This attribute is required, and points to a Internationalized Resource Identifier (IRI).

Another two important elements (which do not appear in the example) are “source” and “rights”. They define the legal issues and rights behind the feed.

One of the advantages of Atom format is that it can be extended. Also this version is accompanied by a protocol [26] for publishing and editing Web Resources. For all its features (security, extensibility, modularization, protocol, etc.), Atom is probably to become the universal standard for reading and writing information on the Web.

However, it is still far from that level. This is our impression after reviewing that most online newspapers publish their content using RSS.

The first atom version was 0.3. Now this is disregarded and replaced by the version 1.0. Atom 1.0 has many of elements and features not mentioned in this document. They can be found in the original reference, see [25].

2.4.2 RSS

The next most popular Web feed format is known as Really Simple Syndication (RSS). It was the first Web feed designed to "feed" Web content. After all the modifications to previous versions 2.0.x is the one currently in use and undergoing progressing research. The last RSS published version is 2.0.11. The peculiarity of version 2.0 is that for first time namespaces were added to a RSS document. Also, new elements adding information [Ayers and Watt 2005], especially for authoring person and software feed's. As in earlier versions, this one has a top level element named "rss" (bolds in Figure 2) . This item has nested an unique "channel"element (bolds and italics) . The channel is the metadata of the feed. All the rest of the elements are nested inside the channel. Some of these children elements are required and others are optional**.**

The RSS requires only three elements in the channel to be working and valid: title, link and description (Figure 2). The flexibility afforded by the optional elements can extend the feed enhancing its functionality .

```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
   <channel>
      <title>Here comes the title</title>
      <link>http://example.com</link>
     <description> Here we describe the feed</description>
   </channel>
</rss>
```

Figure 2: RSS 2.0 document.

Inside a channel, an optional element named "item" can be included. There may be more than one item element per channel. An item may contain children elements such as: title, link, description, publication date and source (Figure 3). All elements of an item are optional but at least one of title or description must be present (for complete reference see **[28]**). The item element serves to represent the metadata of an RSS.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rss xmlns:dc="http://purl.org/dc/elements/1.1/" version="2.0">
  <channel>
    <title>PDC Disaster News</title>
    <link>http://www.pdc.org/</link>
    <description>Aggregated World Disaster News Feed</description>
    <pubDate>Thu, 31 Dec 2009 18:43:35 GMT</pubDate>
    <dc:creator>Pacific Disaster Center</dc:creator>
    <dc:date>2009-12-31T18:43:35Z</dc:date>
  <item>
     <title>PDC World Hazard Briefs - Weather</title>
     <link>http://www.pdc.org/iweb/whb.do?action=weather</link>
    <description>PDC World Hazard Briefs - Weather</description>
    <category>World</category>
     <pubDate>Thu, 31 Dec 2009 18:43:35 GMT</pubDate>
     <guid>http://www.pdc.org/iweb/whb.do?action=weather</guid>
     <dc:date>2009-12-31T18:43:35Z</dc:date>
  </item>
  <item>
    <title>PDC World Hazard Briefs - Other Hazard</title>
    <link>http://www.pdc.org/iweb/whb.do?action=other</link>
    <description>PDC World Hazard Briefs - Other Hazard</description>
    <category>World</category>
    <pubDate>Thu, 31 Dec 2009 18:43:35 GMT</pubDate>
    <guid>http://www.pdc.org/iweb/whb.do?action=other</guid>
    <dc:date>2009-12-31T18:43:35Z</dc:date>
  </item>
</channel>
</rss>
```

Figure 3: Extended RSS 2.0 document.

A summary of the RSS elements is listed below. All of them lie in the channel but may also be inside the item element.

- Description (<description>): Summary of full text of the story
- Link (<link>): Should contain an Universal resource Locator (URL). For channel is the corresponding source. For the item is the URL where points to.
- Title (<title>): Title of the feed. Usually in newspapers, the title of the item is equivalent to the news headline.
- Category (<category>): Is the class, which the feed belong to (i.e, media, science, music, news, academia, etc.).
- Cloud (<cloud>): Is used to manage feeds through remote services using the RssCloud application programming interface (API).

- Copyright (<copyright>): Feed's copyright statement . If the statement does not appear, it should not be assumed that the feed is public.
- Generator (<generator>): Software used to create the feed.
- Image (<image>): Feed's image.
- Language (<language>): Identifies the feed's language based on the W3C for HTML specification. For example, Bolivian Spanish is "es-bo".
- Publication date and time of the feed (<pubDate>): In the format: Sun, 10 Nov 2008 06:30:00 GMT.
- Skip (<skipDays> and <skipHours>) : Indicates the date (day and hour) when the feed is not updated.

2.4.3 Atom and RSS differences

The feed's user community is discussing, which of the two formats is the best. Here, we highlight some differences, most of them taken and slightly modified from [29].

Atom 1.0 is described in the request for comments (RFC) document 4287 [25] (last updated: December 2005), approved by the Internet Engineering Task Force (IETF). RSS 2.0 specification is copyrighted by Harvard University [28] (last updated: July 2003). The other main issues are described below.

- In regard to feed publishing protocol, the Atompub working group is in the late stages of developing the Atom Publishing Protocol [26], closely integrated with the Atom feed format. For its part, RSS can uses MetaWeblog and Blogger APIs based on the XML-RPC protocol [27].
- The required content is: for Atom 1.0 both, feeds and entries should include title, unique identifier and a last-updated time-stamp. While, RSS 2.0 requires feed-level title, link, and description and does not require any other elements inside the item. With respect to the payload issue, Atom has a well-designed payload container. Content must be explicitly labeled as one of: plain text, escaped HTML, well-formed XHTML markup, some other XML vocabulary, base64-encoded binary content or a pointer to Web content not included in the feed. RSS 2.0 may contain either plain text or escaped HTML, but there is no way to indicate which of the two is provided. The RSS 2.0 content model does not permit well-formed XML markup.
- Regarding auto-discovery, Atom has an application/atom+xml Multipurpose Internet Mail Extensions (MIME) type registered with the Internet Assigned Numbers Authority (IANA).

- On extraction and Aggregation: Atom 1.0 allows standalone Atom Entry documents; these could be transferred using any network protocol, for example XMPP. Atom also has support for aggregated feeds, allowing entries to point back to the feed they came from when they are included into other feeds. The only recognized form of RSS 2.0 is an <rss> document.
- About extensibility: Atom 1.0 is in an XML namespace and may contain elements or attributes from other XML namespaces. RSS 2.0 is not in an XML namespace but may contain elements from other XML namespaces.
- URIs: Atom 1.0 specifies use of the XML's built-in xml:base attribute for allowing the use of relative references RSS 2.0 does not specify the handling of relative Uniform Resource Identifier (URI) references. Different feed readers implement differing heuristics for their interpretation. There is no interoperability.
- Software Libraries: Atom 1.0: XML::Atom, XML::Atom::Syndication, FeedParser, Rome, Apache Abdera. RSS 2.0: FeedParser, Rome.
- Language Tagging: Atom uses XML's built-in xml:lang attribute. RSS 2.0 has its own <language> element
- Authors: Both provide the option to specify the authorship.
- Schema: Atom 1.0 includes a (non-normative) ISO-Standard RelaxNG schema, to support those who want to check the validity of data advertised as Atom 1.0. The RSS 2.0 specification does not includes schema.

2.4.4 GeoFeed examples

Usually a Web sites does not provide Web feeds including coded geographic information. Thus, when feeds are geotagged they change to become GeoRSS or GeoAtom, here both are named GeoFeeds. There are two GeoFeed formats one simple or lightweight, and another fully featured. The later is a profile of the Geography Markup Language (GML).

The light version requires one tag per object geometry to be valid. And also include the GeoRSS namespace declared in the XML document (Figure 4).

```
<?xml version="1.0" encoding="utf-8"?>
  <feed xmlns="http://www.w3.org/2005/Atom"
xmlns:georss="http://www.georss.org/georss">
      <title>Earthquakes</title>
      <subtitle>International earthquake observation labs</subtitle>
      <link href="http://example.org/"/>
      <updated>2005-12-13T18:30:02Z</updated>
      <author>
          <name>Dr. Thaddeus Remor</name>
          <email>tremor@quakelab.edu</email>
      </author>
      <id>urn:uuid:60a76c80-d399-11d9-b93C-0003939e0af6</id>
  <entry>
      <title>M 3.2, Mona Passage</title>
      <link href="http://example.org/2005/09/09/atom01"/>
      <id>urn:uuid:1225c695-cfb8-4ebb-aaaa-80da344efa6a</id>
      <updated>2005-08-17T07:02:32Z</updated>
      <summary>We just had a big one.</summary>
      <georss:point>45.256 -71.92</georss:point>
  </entry>
  </feed>
```

Figure 4: Modified from GeoRSS [15]

Web examples working with GeoFeeds can be seen in [30], [31] and [7]. Figure 5 refers to the second, which is committed to fetching news feeds from the BBC-UK edition, and display them in a map.

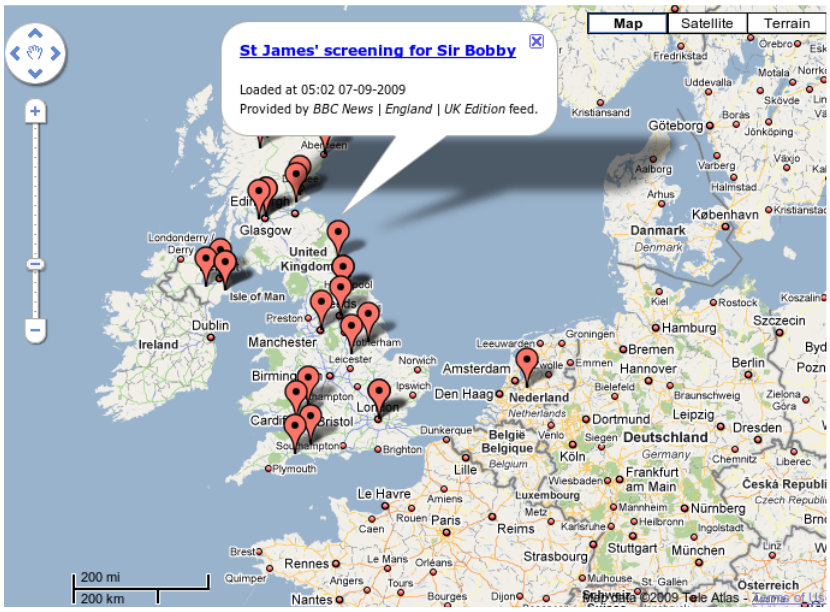

Figure 5: BBC news feeds for UK on Google Maps

The [7] example is well linked with the Geonames project [Wick and Becker 2007].

This initiative has taken several approaches to geodata gathering and dissemination. One of the most important factors that make this project successful that it's supported by a vast geographical gazetteer. Besides their many names of different geographical features, at least 2.5 million belong to a placenames. The database is free to download and the Geonames community has added several services to access it. For example, any user can search for names of places and geographical features, view on a map, send by email and export them in different formats. Registered users can edit the database, for example adding new records. That means the database is under constant review and is continuously updated. The project also offers an API to employ their extensibility.

The Geonames team has developed a service able to convert RSS to GeoRSS. The translator supports different languages. So that the processor work which needs well formed input files and valid XML. A summary of how the parser works is described below:

The tokenizer algorithm they introduced extracts the text contained in the title and description of each item and combines them in a single text. Star ting with the description, the processor tokenizes the word handing over to a par t of speech tagger for further processing using word frequency techniques. Afterward, a query generator discards irrelevant tokens and weighs them to be run against to the large Geonames database in different ways. After this, the algorithm computes a second score for the most relevant toponyms. The last processing routine gets help from feed's items names to improve the disambiguation of the placenames. A test running by Geonames, suggest 90 % accuracy in extracting the correct terms. [Wick and Becker 2007] and [14].

We tested the service by entering the link from a Mexican newspaper. The news broadcast by this provider is in Spanish and is divided into sections. In our test we used the section "states". In this section, nearly all titles include one or more names of some state of the Mexican Republic. The results of the parser were accurate when the title included only a name. However, when there was more than one, the parser took the first and assigned the coordinates (latitude, longitude) assuming that the news only corresponded to that place. We believe Geonames' parser is practical when users want to discover at a glance where news is happening, without being concerned with the precision of the place in cases where there is more than one.

Metacarta ([44]) offers Web mapping applications to process text within news stories using proprietary technology. It extracts geographic places mentioned in new stories and identifies their respective latitude and longitude coordinates. The places are then represented through a customized geobrowser. The company offers a free API to test many other related services.

On the other hand, looking at Web, we detected other parser dealing with feed and GeoFeeds. **GeoFeed** [21] is a beta processing tool for identifying and extracting geographical references from feeds. This tool returns a news feed including a tag with the geographical places within the original feeds. This product offers a “loose API” and is only available for detecting cities and States of U.S.A. The developer company plans to provide better toponym and disambiguation processing to improve identification accuracy. At the writing of this work, their page does not provide compressible explanation of the parsing algorithm. A resulting example seems to rank the cities according to the possible States they belong. Alternatively, we explored displaying a map from their “Recent Feed” Web link to check the location of the returning results, but got an error even after attempting it several times.

3 Methodology

In this section we present the conceptual design of our extraction approach. This aims to extract geo-information from unstructured text to populate a database. We begin introducing the methodology with the architecture of the system, followed by a description of the data model, including a comprehensive explanation of its components. Afterward, we describe the steps that fulfill the experiment.

3.1 Architecture

The architecture presented here is set up under the client-server design. Our system include four components (Figure 6). The first one consist of an agent stylized for reading Web feeds from Internet sources continuously. Those feeds are accessed by their universal resource locator (URL) and used as input to be analyzed in the second component. This component is meant to filter against a built-in catalog and extract geographic information from those feeds producing an output (third component). The output is turned into a database, which will be ready to be accessed by a geoclient (fourth component).

The Web feeds alike the geoclient are on the client side. The catalogs, the parser processor and the database rest on the server side. All the components get communicated on Internet through a common gateway interface (CGI). A CGI is a interface for external programs to interact with information servers such as Web servers [42]. The interaction between programs in the CGI can be done over different business layers.

There are many technologies involved in handling the data transactions and processes phases. The code development is constructed using the philosophy of soft-

ware ecology, which allows use components from several sources and be reused to create new ones (Cook and Daniels 1994).

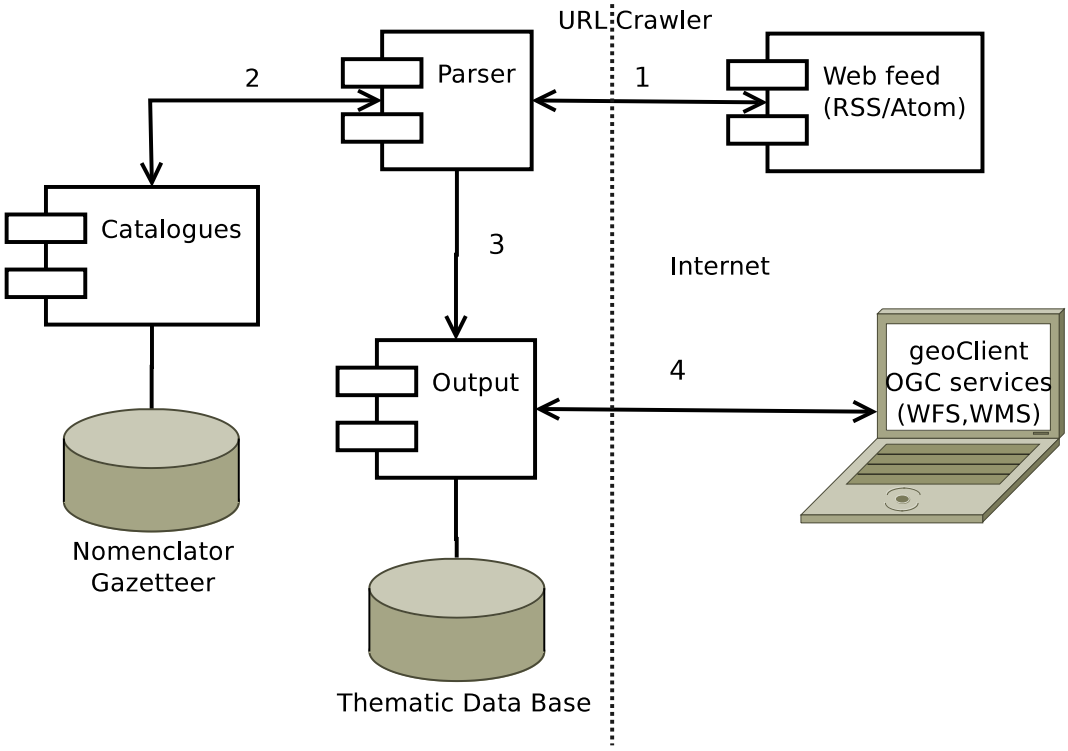

Figure 6: System architecture

3.2 Data Model

In this section we describe the data model used for our system. The data model gathers the information required in situations such as those described in section 1.2. Therefore, it includes places, attributes related to risk and disaster subject, and additional layers.

The data model is divided into three parts (Figure 7). One concerning the *national geostatistical framework*. Secondly, the *risk framework* package, and thirdly a *thematic* package containing complementary layers. The constraints of a news story is stored in the database described in section 3.3.

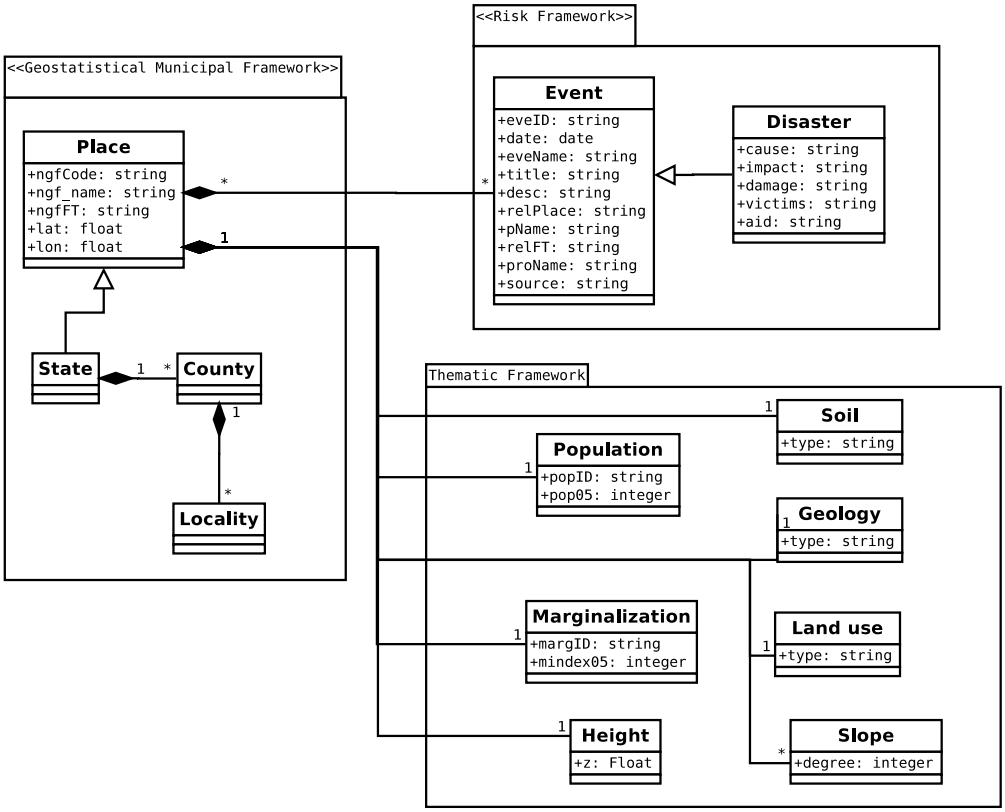

Figure 7: Data model

National geostatistical framework

The national geostatistical framework (NGF) package has the classes Place, State, County (county is synonym of municipality) and Locality (Figure 7). At the same time State class is composed by County and County by Locality. Since Place is the generic class, the remaining classes are its children and inherit its features. The attributes of the Place class are: name, code, type and geographic coordinates. Those features are modeled following the administrative structure used for the Mexican territory described below.

The NGF is a system created, maintained and distributed by the Mexican national mapping agency: National Institute of Statistics, Geography and Informatics (INEGI, initials in Spanish [23]). The NGF was designed to correctly reference the statistical information from census and surveys at their corresponding geographic location. The catalog used in our work is based on the NGF structure, which is divided in three levels: state, municipality and locality. This nomenclator contains most of the Mexican placenames.

The NGF dictionary contains 31 names of each state and one for the Federal District. Unlike states, the Federal District is considered different by the legislation. But here we treat it as a state because our processor handles its code in the same way.

The 32 states are depicted in Figure 8.b. INEGI uses two digits to code state entities. The range starts from 01 to 32. Those numbers are assigned consecutively based on alphabetic sorting. Therefore, “Aguascalientes” is heading the list and “Zacatecas” comes last.

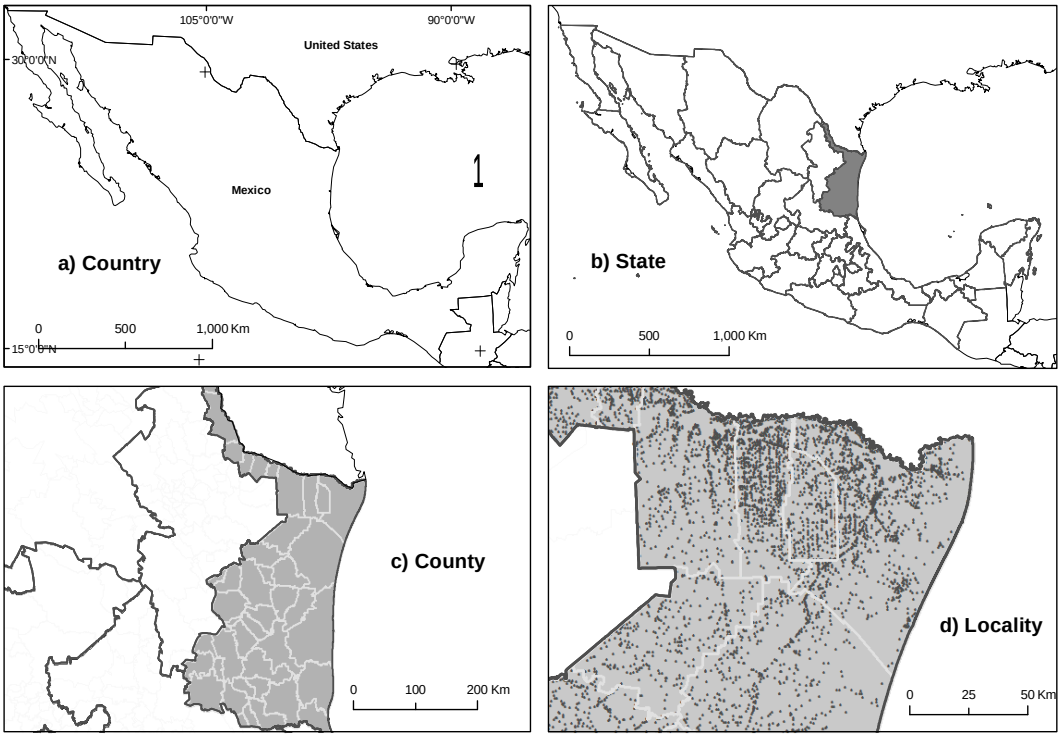

Figure 8: Administrative divisions in Mexico. Source: [23]

States are composed of one or more municipalities. That composition can be seen in Figures 7 and 8.c. The national mapping agency assigned three-digit numbers to code municipalities. The chain of municipalities ranges from 001 to 999 (Table 1). As of September 2009 [24], the catalog has a total of 2456 municipalities.

The next level in the administrative hierarchy below counties are localities. Those entities are defined by places having one or more houses, which can be inhabited or uninhabited. The name of localities are given by their local people and if not they are officially set by law. By November 2009 INEGI recorded around 293,716 localities [24]. Most of them (284,741) were located in rural areas. INEGI uses four digits to encode localities (Table 1). Some of these villages spreading on the state of Tamaulipas are depicted in Figure 8.d.

In summary, the geostatistical keys are set depending on their disaggregation level. The order that their code follows is: SS +CCC+ LLLL. Where: SS = state, CCC= county and LLL= Locality. Table 1 provides the list of the NGF standard and gives an example .

Table 1: NGF code forming.

Name	Code		Name	Range
State	00	28	Tamaulipas	28:28
County	000	005	Burgos	28001: 28043
Locality	0000	0052	Las Flores	280010001: 280430510
SS+CCC+LLLL	000000000	280050052	Tamaulipas/ Burgos/ Las Flores	No apply

For instance, the state entity listed in Table 1 and grayed on Figure 8 (northeast corner) corresponds to Tamaulipas. Which is the 28th state in the country, comprising 43 counties and 520 localities.

The reference system for the NGF entities are expressed in geographic coordinates (lat / lon) [23]. States and counties are originally provided by INEGI as polygons (2D), but it should be stressed that in our database, they will be stored as points (0D), using the centroids from the former features.

Table 2: Place table

Field	Description
ngfCode	NGF Code. State 00; County 000, Locality 0000
ngfName	Placename in the NGF
ngfFT	Feature Type (State, County, Locality) in the NGF
lat	Latitude coordinate
lon	Longitude coordinate

Risk Framework

We categorize the Risk Framework package into two classes, events and disasters. The first is generic and includes the minimum attributes that the processor could extract from a story related to a natural hazard. The Event class exists if at least one place from the NGF is documented in the news. This indicates that there may be a relationship of one place to many events. If so a record is generated for every location. The same occurs if many places and one event is described or multiple locations match multiple events.

Similarly, when the processor identifies specific categories related to a disaster, then includes the class Disaster. This in turn inherits the attributes of the class Event. The design of the two classes is complementary, allowing us to maintain control and verify the precision in the outcome database.

The metadata for the attributes of Event and Disaster tables are set in Figures 3 a 4 consecutively.

Table 3: Event table

Name	Description
eventID	Is the concatenation of the NGF code (ngfCode) from Place table and date attribute of the Event table.
date	Publishing date of the news on the feed sourced.
eveName	Name of the event (rain, landslide, earthquake, flood, etc.).
title	Title of the news story (title element).
desc	Text in the *description* element of the news feed story.
relPlace	Related places found in the news.
relFT	Feature types found in the news (e.g. river, range, park, etc.)
proName	Proper names found in the news.
Source	Link to the news story (URL). Links between the extracted information and the original documents are maintained to allow the user to reference context ([40]).

Table 4: Disaster table

Name	Description
cause	I.e., if a flood was triggered by heavy rains, or by a hurricane. Other causes of disasters are poverty, weakness on the physical infrastructure and social preparation.
impact	Is the sum of victim and damage.
damage	Economic impact reported in a given currency (i.e., US$, €).
victim	Number or reported evacuations, deaths, homeless. Other damaged objects (i.e., bridges broken/collapsed; acres/hectares flooded; houses destroyed; etc.).
aid	Name of a potential aid supplier (i.e., United Nations, Red Cross, Civil Protection).

Most of the constraints in handling issues concerning places and attributes to fill in the database are explained in section 3.3.

Thematic Framework

We included a package named Thematic Framework because Mexican territory extends roughly over two million square kilometers. Such extension exposes many places to multiple hazards and we think thematic variables may help to reveal social, economic and physical vulnerability across that space.

This package comprises social and environmental variables. On the social side, the package contains the total population [23] and the marginalization index from

the National Council of Population for Mexico (CONAPO, acronym in Spanish) [33]. These indicators are available for the NGF division. The population data is obtained from the national census every ten years. Recently a population counting has been implemented every five years since 1995. The marginalization index follows the same NGF structure and also is provided every five years since 1995 [33].

The environmental indicators we included are: soil type, geology, land use, slope steepness and height. Unlike the previous discrete indicators, these are continuous fields. They can be included by performing spatial operations (such as spatial joint) on the TDB points. Some characteristic of the layers are described below.

Table 5: Thematic framework

Table	**Content**	**Description** (sample)	**Source/ resolution**
Population	Relative population	Census populations: every ten years. Count estimations every five.	INEGI/ NGF
Marginalization	Marg. index	Index comprising several social aspects of social exclusion.	CONAPO/ NGF
Height	Z values	Height value above sea level	SRTM/ 30mpp
Soil	Soil type	Aridisols, Entisols, Gelisols, etc.	INEGI/ 1:250k
Geology	Geology type	Bed rock, Dolomite, etc.	INEGI/ 1:250k 1:50k
Land use	Land use	Urban area, vegetation, (mangrove, forest land), field, crops, etc.	SEMARNAT/ 1:250k
Slope	Slope steepness/integer	Degree of slope inclination	Derived from SRTM

Height and slope layers can be reaped from the SRTM (Shuttle Radar Topography Mission [34]). Soil and geology layers may be sourced from INEGI [23] at scales of one to one million and those listed in Table 5. In Mexico Land Use data is collected by the national and sub-national environmental agencies. The cartographic scale and time frequency of this layer varies.

Complementary information would provide insight into the time of data access and visualization. It would also give a better overview and understanding of the actual risk conditions where TDB spots are recorded. There are many more layers that can be added and combined with the TBA. Here we include only some that we considered relevant.

3.3 News analysis

In this section we explain the methods to extract geographic information from Web feeds. First we highlight some of the news features, and show how text data contained in digital news papers can fit our problem. Next, we explain the main characteristics of each of the processing phases.

A word on digital newspapers

We assume that texts on Internet newspapers contain spatial, temporal and thematic information. For the first two types are well established reference system. Despite locations and their coordinates are defined in gazetteers, the attributes referring to them do not. The same for the name of places (placenames) because we refer to them in multiple ways, and they change over time .

Since semantic reference systems, yet is an open research topic [Kuhn 2003], our biggest challenge is to extract the attributes to their corresponding location. And then generate the records presented in the data model (Figure 7). This semantic problem has been long sought since information extractions begun [40].

Regarding news broadcasters, they emit their feeds either in RSS or Atom. The information in digital newspapers can usually be dived in topic sections or as complete edition. Therefore, news feeds are split or keep as full document containing all the stories for a daily edition. The BBC for instance, offers news feeds separately fo*r topics such as world zones*, *health, science, technology* and *entertainment* to mention a few. That company also offers the complete edition in a single feed[20]. This may brings drawbacks when a text processor deals with the mixed information contained in Web feeds. Because it may "confuse" the disambiguation processor at the time to identify categories our interest among a pull of meanings.

In the other hand, there are some advantages that newspapers offer. We consider highly important the writing style used by journalist when they report news stories. Usually they use capitalization on proper names and assign administrative hierarchies when documenting locations. For instance, in Figure 9, Europe appears in the title and a list of other European countries remains in the description element. Following that hierarchy, if an agent aims for retrieving the countries inside Europe, the processor may look through a gazetteer to restrict the search under the geographical extent of that continent.

Another advantage for extracting information from news is that each story is inside an item in the description element. Most news feeds contain only an excerpt or the first paragraph from the complete story. Besides, items have their own date and link attributes (see Figure 9 and section 2.4), this fact make less complex to identify the attribute data for each story.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<rss version="2.0" xmlns:media="http://search.yahoo.com/mrss/">
  <channel>
    <title>BBC News | Europe | World Edition</title>

    <link>http://news.bbc.co.uk/go/rss/-/2/hi/europe/default.stm</link>
    <description>Get the latest BBC News from Europe: headlines, features and analysis from BBC
correspondents across the European Union, EU, and the rest of Europe.</description>
    <language>en-gb</language>
    <lastBuildDate>Thu, 07 Jan 2010 12:37:43 GMT</lastBuildDate>
    <copyright>Copyright: (C) British Broadcasting Corporation, see
http://news.bbc.co.uk/2/hi/help/rss/4498287.stm for terms and conditions of reuse</copyright>
    <docs>http://www.bbc.co.uk/syndication/</docs><ttl>15</ttl><image>
    <title>BBC News</title>
    <url>http://news.bbc.co.uk/nol/shared/img/bbc_news_120x60.gif</url>
    <link>http://news.bbc.co.uk/go/rss/-/2/hi/europe
/default.stm</link></image>
  <item>
    <title>Airport chaos as Europe freezes</title>
    <description>The icy weather gripping northern Europe disrupts flights in the
UK, France, the Irish Republic and the Netherlands.</description>

    <link>http://news.bbc.co.uk/go/rss/-/2/hi/europe/8445613.stm</link>
    <guid isPermaLink="false">http://news.bbc.co.uk/1/hi/world/europe/8445613.stm</guid>
    <pubDate>Thu, 07 Jan 2010 12:03:33 GMT</pubDate>
    <category>Europe</category>
  </item>

</channel>
</rss>
```

Figure 9: BBC news feed sample

Web feed in Figure 9 shows a short news story. On it can be straightforward to extract the time because is explicit but attributes (and placenames as part of) are implicit.

In the same example, it can be quite simple to recognize that the icy weather disrupts flights in four countries. But what if the story further says: “Canceled flights in the UK have exceeded 50 percent of normal” to which of the four mentioned countries do we must allocated the percentage of flights canceled? Obviously, from the context if one person reads this text would designate the attribute to the UK, but for an automatic processor can be more complex determine which places attributes belong to. The problem can be summarized as: how to know which of the dynamic placenames point to static places, and which from the heterogeneous attributes corresponds to the right placename.

Table 6 summarizes the the three types of information that can be sourced from Web Feeds.

Table 6: Information types

Info type	**Source**	Example from Fig. 9.
Places (where?)	Gazetteers	Coordinates (x,y) for Europe, UK, France, Ireland and the Netherlands.
Time (when?)	News feeds	07 Jan 2010 GMT
Attributes (place-names/what/how? etc.)	News feeds	Weather gripping, disrupts flights, airports Europe, UK, France, Ireland and the Netherlands

Processing phases

The processing of news content is divided into two phases. The first begins using an analyzer that utilizes NER techniques to identify information related to our domain. If this step does not success at all, a second processing phase is carried out (Figure 10).

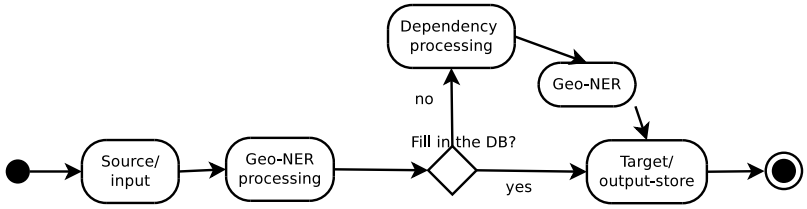

Figure 10: Process work-flow

The first step before processing a feed is fetching it. This can be done using a program supporting HTTP transactions by accessing to an URL of a newspaper. Despite of that, for this experiment we use a news feed create for test bed purposes. The document is about of a hundred news stories. All of them come from the same newspaper [32] covering press releases for whole the Mexican Republic. The document contains news of different dates and topics. And from different sections of the newspaper (National, International, States, Politics, etc.). There are 196 stories, around 60 (30%) of them are relevant for the disaster domain; the 136 (60%) remaining are senseless to this topic. We deiced including more irrelevant stories than relevant to test how the disambiguation methods perform.

Story in Figure 11 is included in the corpus and used to illustrate some parts of the processing stages. This news comes originally in Spanish language but we give parallel explanation in English (bold tags).

<?xml version="1.0"?>
<rss
xmlns:content="http://purl.org/rss/1.0/modules/content/" version="2.0">
<channel>

<title>La Jornada</title>
<pubDate>Wed, 17 Sep 2008 09:05:42 GMT</pubDate>
<description>Noticias del diario mexicano La Jornada</description>
<link>http://www.jornada.unam.mx/2009/09/17/</link>
<Item>

<title>Provoca Ike inundaciones y tres muertes en Nuevo León</title>
<title> *Ike causes flooding and three deaths in Nuevo Leon* **</title>**

<link>http://www.jornada.unam.mx/2008/09/17/index.php?
section=estados&article=039n1est&partner=rss</link>
<pubDate>Wed, 17 Sep 2008 09:05:42 GMT**</pubDate>**

<description>Las lluvias que provocó el huracán Ike causaron la muerte de tres personas en la sierra del municipio de Santiago, Nuevo León, y desbordaron la presa El Cuchillo, lo que inundó durante más de 24 horas algunos tramos de las carreteras que van de Monterrey a Reynosa y a Ciudad Victoria, y de ésta a San Luis Potosí. En Reynosa, el oleaje arrasó con al menos 100 metros de la barda de contención de playa Bagdad y devastó el área de palapas.</description>

<description>The rains caused by Hurricane Ike killed three people in the mountains of the municipality of Santiago, Nuevo Leon, and overflowed El Cuchillo dam, which flooded more than 24 hours some parts of the roads that come from Monterrey to Reynosa and Ciudad Victoria, and from there to San Luis Potosi. In Reynosa, the waves swept at least 100 meters from the containment fence in Baghdad beach and devastated the palm shelter area.**</description>**
</item>
</channel>
</rss>

Figure 11: Sample news story

3.3.1 Geo-NER

In this point we explain the main features of the text processor. Also we describe how it handles Web feeds to identify semantic content related to our topic. At the end we mention the characteristic of the output and how is further filtered.

The main goal our processor is to identify information relied on Web feeds semantically related to disaster an risk domains. We built a tool named Geo-NER to do the task. It uses name entity recognition functions to create text categories. Geo-NER is founded in heuristics expressed in several components such as preconfigured

sets of regular expressions; stop-word lists; the NGF toponym catalog; thesaurus about terms related to natural phenomena; and other pattern based grammars to match entities of interest.

Regular expressions (RE) specify a set of strings that match it. From RE we can form other RE, which makes quite simple to create complex RE from single ones.

Our parse is developed in SAX (Simple API for XML) to access elements on XML files, and uses *Python RE* and *String* libraries to cope with text identification of thematic words. For instance, RE may help us to find capitalized terms and words related to events and causes as those shown in Table 7.

Table 7: Example of regular expressions

Focus	**Spanish**	**English**
Event	|[Ll]luvias?	|[Rr]rain?
Cause	[Dd]eslaves?	[Ll]andslides?

Inside the RE is included a series of categories to of the disaster domain based on the data model. For instance, configured patterns of grammars matching natural phenomena that may become dangerous, e.g: rain, heavy-rains, tropical storms, hurricanes and earthquakes. Other chain of strings in RE are associated to terms related to victims caused in disasters. For example: dead, affected people, damage. Likewise, we provide many organization names for NGOs, aid bodies and government agencies (Civil Protection). Most of the terms included are collected after a theoretical review on risk/disaster aspects from [1] and [38].

Also we included a stopwords list, which prevents the parser to to look for words that does not provide clues for identify targeting words.

The processor also encompasses access to NGF gazeteer,in which the processor look up to perform geo and non-geo disambiguation. Re-casting, this dictionary includes toponyms and codes for 32 states, about 2,500 municipalities and 200,000 villages.

Geo-NER uses the above components to identify entities of our interest. After obtaining the feed of a given URL, the feed is stored as an object in memory and a parser points to it. Then, the algorithm in Figure 1 iterates over the text on those feeds. It starts with the disambiguation of placenames followed by the attribute identification. The complete algorithm is described in Appendix: A.

Algorithm 1 Geo-disambiguation by codes

1. The processor runs over the feed title and description to identify placenames in the text.
2. Compare those names against the NGF catalog.
3. If found toponyms, the candidates along their codes are stored, otherwise ends.
4. After identifying placenames, the program starts cheking top (state) to bottom (localities) for codes of the administrative division, holding their ranges. Those codes will restrict a new iteration seeking municipality names over the text based only within the range.
5. If found toponyms at one lower administrative level go to next step, otherwise ends.
6. The previous iteration is repeated for municipalities to check for localities.

The above steps are executed until the parse return the maximum geo-disambiguation in terms of administrative level. That is, if the parser can not find names of municipalities under states stops storing just states. But if in the process municipalities are found they are stored and used to check if there is placenames of localities. The resulting placenames are attempted to be at the finest administrative level, which corresponds to localities. Consequently, from the same story we could generate records at different administrative levels.

After looking for placenames, the program follows seeking in the text for categories of attributes related to risks and disasters. Geo-NER uses the predefined RE and the other lists to perform that action. Most of non-geo disambiguation is eliminated implicitly because the parser only uses toponyms stored in the NGF gazetteer. Filtering placenames reported in news depends upon if they exist in the NGF gazetteer.

Output

When the processor finishes returns a list of attributes, which may or may not contain, the complete list of attributes in the tables of the data model. A output sample looks like Figure 12.

The list that the output returns match with each block of text used by the processor. That means in may be more than one block per paragraph, usually divided by do. There are common attributes for all text blocks; they are within the news item, including: link, date and title. Despite the title is also a block of text, is excluded of further analysis because is a common attribute.

Once the list is returned, is analyzed to verify whether accomplishes some constraints (explained later on) to make them fill or not the data base. For example,

if in a news story there are two blocks of text, there are three possibilities to post-process them: either both are stored in the database; only one is dumped to the data base and the other is taken to the next step; or both are set out to be reanalyzed. The same occurs if there are more than two blocks.

From Web feed in Figure 11
Block 1) The rains caused by Hurricane Ike killed three people in the mountains of the municipality of Santiago, Nuevo Leon, and overflowed El Cuchillo dam, which flooded more than 24 hours some parts of the roads that come from Monterrey to Reynosa and Ciudad Victoria, and from there to San Luis Potosi.
————————-Placenames————————————-
found locality: Santiago [19049001]
found locality: San Luis Potosí [240280001]
————————-Attributes————————————
['Cause', 'lluvias'] /*rains*
['Cause', 'huracán'] /*hurricane*
['victim', 'muerte'] /*deaths*
['relatedPlace', 'en la sierra del municipio de Santiago'] /*mountains of the municipality of Santiago*
['ngfFeatureType', 'municipio']/*municipality*
['relatedFT', 'presa'] /*dam*
['eveName', 'inundó'] /***flooded***
['relatedFT', 'carreteras']/*roads*

From Web feed in Figure 11

Block 2) In Reynosa, the waves swept at least 100 meters from the containment fence in Baghdad beach and devastated the palm shelter area.
————————-Placenames————————————-
found municipality : Reynosa [28032]
————————-Attributes————————————

Common Web feed elements (Figure 11) for both bolck of text above.
title-> Ike causes flooding and three deaths in Nuevo Leon
pubdate->Wed, 17 Sep 2008 09:05:42 GMT
link->http://www.jornada.unam.mx/2008/09/17/index.php?section=estados&article=039n1est&partner=rss

Figure 12: Geo-NER sample output

A filter is run over each list to make the decision whether a text block fills a table or not. The constraints are: only one placename must be considered at the finest level of granularity of the NGF; at least one of the output attributes must report an

event or *cause* of a *risky* ocurrence. The steps carried out by the processor to filter the outputs are explained in Appendix A.

Because of the mentioned constraints, Block 1 (Figure 12) contains more than one place in the same administrative level, thus, it needs to be passed to the next processing phase, which carry a grammatical analysis through the block.

The Block 2 (Figure 12) is excluded from any further processing because it does not have any of the required attributes (event name or cause).

The Appendix A shows the blocks of a news story that accomplished the constraints, and their attributes that are entered to the database. In the same section the process to achieve the fill in of the database is summarized. The process is basically to parse the output list and create tuples that match the structure of our data model; and after that other agent fill the database. The TDB population begins with the Event table, which in turn can be easily joined with table *Place* using the NGF codes. Depending on the attributes of the output list returned by Geo-NER process, the table Event may include attributes for Disaster table too. The Appendix A shows an example of three tables (Place, Event, Disaster) merged in one.

Text blocks that need to be treated under grammar analysis, are passed as plain blocks of text. The way how they are treated is explained in the following section

3.3.2 Dependency grammar analysis

In this section we present dependency-grammar analysis for text blocks of Web feeds.

Dependency grammar analysis (DGA) aids to identify and assign attributes of interest to their corresponding placenames appearing in news stories. To do this, DGA tries to identify relationships between words through sentences based in the grammatical rules of a specific language; in this case Spanish. We use a processor ("analyzer") included in Freeling, a language library that provides language analysis services [Carrera et al. 2008]. Freeling handles around seven languages including Spanish and English. It is also supported by several knowledge dictionaries with encoded semantic information. Freeling is able to handle multi word detection,number, dates, and quantities recognition (see [36]).

The analyzer is able to processes blocks of texts and compute their grammatical relationships; after that it returns a dependency tree like those in Figures 13 and 14. The first depicts the tree analyzed for the block text from the Web feed in Figure11. The second shows the tree as an image from a sentence corresponding to the title in the same feed.

```
top(and and
    [(caused cause
        [ (rains rain [(The the ] (by by [(killed kill
            [(Hurricane_Ike hurricane_ike (people people [ (three 3 Z ] (in in
                [(mountains mountain [ (the the (of of
                        [ (municipality municipality [ (the the (of of
                            [ (Santiago santiago ]
                        ]
                    ]
                ]
            ]
        ]
    ]
parenth(Nuevo_Leon nuevo_leon
    [ (overflowed overflow [ (and and [ (dam dam
        [ (the the (El_Cuchillo el_cuchillo (which which
            [ (flooded flood [(more more [(than than [(hours hour [(24 24
        ]
    ]
]
(during during [ (stretches stretch [ (some some (of of [(roads road ] ]
](from from [ (Monterrey monterrey ] (to to [(Reynosa reynosa ]
    ]
  ]
](Ciudad_Victoria ciudad_victoria ] ](and and (from from [(there there ](to to
[(San_Luis_Potosi san_luis_potosi]
]
```

Figure 13: Dependency tree of text in 12

The placename in the sentence in Figure 14 is straightforward to disambiguate, because the sentence is short and there is only one placename; but the sentences in Figure 13 are more complex. It is because there is more than one place (in bold) and many attributes. We mention this problem in the previous sections, and our approach to deal with it is to mix Freeling and Geo-NER capabilities.

This processing phase takes the blocks of text that could not be completely disambiguatedin Geo-NER. So, they are those blocks that did not comply with the constraints, to directly fill the database. That is the case of “Block 1” in 12 and its corresponding dependency tree is in Figure 13.

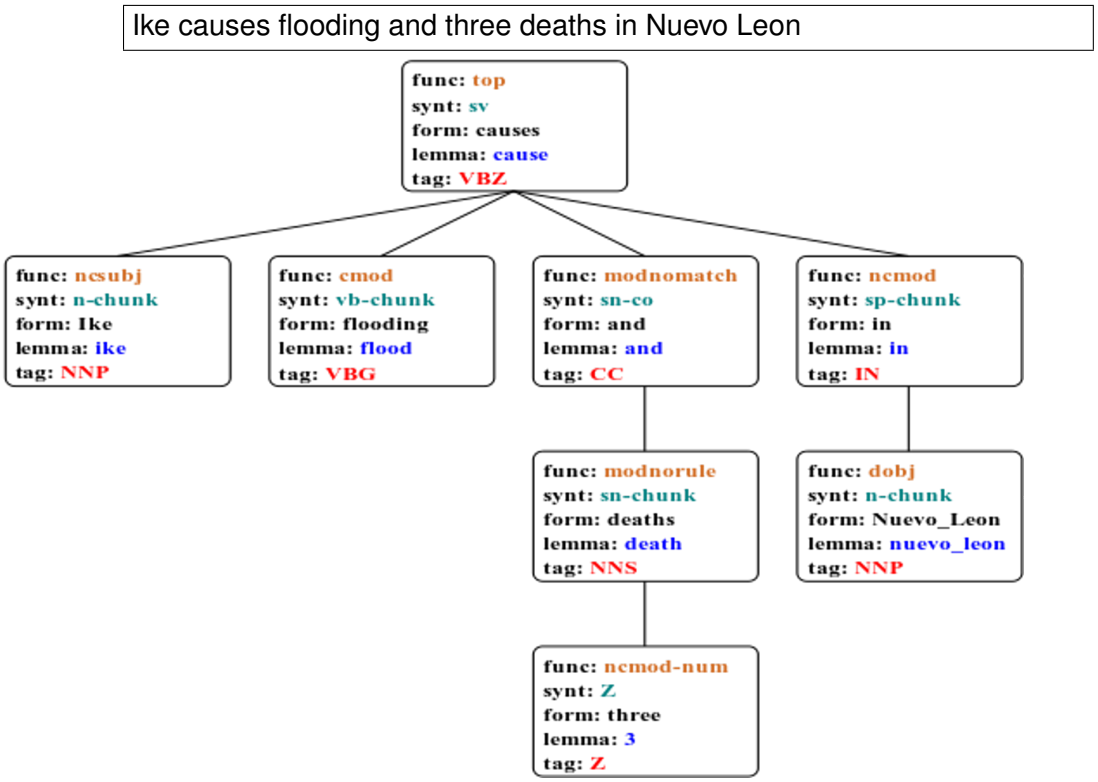

Figure 14: Image of a dependency tree of a single sentence

For each block of text non disambiguated in the previous section a dependency analysis is performed. Therefore, for each a dependency tree is generated. After that, a custom algorithm (2) is carried to process the trees. The algorithm is made upon the assumptions that each tree may have more than one branch; branches may have leaves, which can be attributes and/or placenames. We assume that attributes and placenames are more (or less) related in the context of a news story based in their distance inside and outside the branches.

Algorithm 2 explains the way in which the processor identifies placenames and attributes over the tree. Then, checks their distance based on the depth of the leaves through the branches as a hierarchy (brackets in Figure 13). Therefore, words (leaves) positioned at the left-most part of the branch, are furthest from those words at the right-most part. The same assumption is made between branches from top to bottom.

The first step the processor takes over the text blocks is to look for *events* and *causes;* those attributes usually refer to places. After, the processor checks if places reported in these blocks mach some placenames in the NGF gazetteer, if the processor finds placenames it continues running, otherwise it finishes.

Algorithm 2 Post-processing dependency tree

The program takes the dependency tree to process and convert it into a nested list of leaves.

1. Geo-NER starts identifying in the tree whether there is a cause or event attribute, if there is nothing it ends, otherwise it continues,
2. Geo-NER iterates again crossing the tree looking for placenames that matches with those in our gazetteer, if there is no names the program it ends, otherwise it continues,
3. At this point, other process iterates over the tree again, to convert the nested list into a new list of tuples containing placenames and attributes. The tuples are defined as structured lists according to the database but sorted without ordering (refer to the box below):

   [A] [B] [C][D][...], Where:
   A = Placename, B = Event, C = Cause, D = relPlace, etc.

4. The iteration continues, reviewing the tree from top to bottom and from left to right examining the tuples, and takes the following conditions

   (a) if, within the the first branch there is the event or cause attribute, those attributes are assigned to the placename,

   (b) if there are multiple placenames, the mentioned attribute is assigned to the nearest in the branch. If there are two places (leaves) at the same level, attributes are assigned to the lower placename found in the administrative hierarchy (see disambiguation by NGF codes in point 6 of Appendix A)

   (c) the same routine is performed for the remaining branches storing placenames and attributes found previously in the memory. If more attributes are found, the program stores them along the already detected (event and/or cause). Once the loop finishes, it returns an output list similar to that in the box below:

   Cluster 1.) Place1/Event 1/Cause1/Victim1
   Cluster 2.) ...../Event 2/..../Impact
   Cluster 3.) Place2/..../Cause2/Victim2

   (d) next, another process is run over the list from top to bottom to make a permutation between the tuples. The process stops if it finds a placename; this will ensure each record has unique placename and their corresponding attributes.

   Place1/Event 1, Event 2/Cause1/Victim1/Impact
   Place2/..../Cause2/Victim2

The output list resulted from the algorithm should correspond to the disambiguated placenames and their corresponding attributes. This output is parsed to create new tuples similar to those in Figure 15.

```
————————-Placename selected to assign the attributes————————-
['ngfName','Santiago'], [ngfCode', '190490001']:

————————-Attributes————————-
[['proName', 'Ike '], ['Cause', rains, 'hurricane'], [ 'victim', 'deaths'], ['ngf_Name',
'municipality'], ['relPlace', 'municipality Santiago '], ['relFT', 'dam', 'roads'] ]

————————Common elements for this and other blocks in the same news story————————
['date','Wed, 17 Sep 2008 09:05:42 GMT']
['title', 'Ike causes flooding and three deaths in Nuevo Leon']
['link', 'http://www.jornada.unam.mx/2008/09/17/index.php?section=estados&article
=039n1est&partner=rss']
```

Figure 15: Output list from dependency tree

This list of tuples is parsed again to fill the database. Then, the records stored in the database are ready to be accessed by a geoclient. An example of a record is shown in Table 10 of Appendix A.

After having processed the testing corpus of text, we can evaluate the overall results of the TDB.

4 Results and discussion

Here we analyze the results of the extraction phases. We use the same corpus of news presented in section 3.3 as a benchmark for analyzing the results. The empirical analysis is carried out in two parts: general and particular. In the former we review only the disambiguated news from those non disambiguated. The particular review explains an empirical analysis made over the news disambiguated.

In the general review (Table 8) we counted by hand the number of news items that were disambiguated for the domain context. We named them irrelevant and relevant. Conversely to relevant news items, irrelevant ones are those in which our processor did not find any placename or event/cause attributes. We also counted the number of non disambiguated news items. Similarly, we detected those stories relevant for the risk and disaster domains but that were ignored by our processor.

Table 8: General evaluation

	Irrelevant	Relevant	Relevant but ignored	Total
No of stories	137	55	4	196
%	69	29	2	100

From the news corpus only four stories (2%) were ignored by the processor; the rest (~98%) was properly disambiguated. Around 70 percent of news in the corpus

were not relevant for risk and disaster domains and the remaining (~30%) were relevant. This later set is equivalent to 55 news stories.

For the particular analysis we counted by hand the text blocks inside the 55 relevant news stories. There are 85 blocks, but only 70 became candidates to be records in the database. The 15 blocks excluded were cases that contained neither placenames nor event/causes.

To rank the block quality, first we created categories; after, we identified which processor handled each block and what kind of placenames they produced on each category.

The qualitative categories are five, ranked in ordinal scale from 1-non matching- to 5 -best matching-. Here matching means whether attributes matched the right place, the relevance of the attributes for the disaster domain, and whether the disambiguated placename corresponded to the administrative level of the place described in the news. The definition of each mark is the following:

- 5 = The disambiguated placename matches the attributes and the other way around. It must include attributes from the Event and Disaster tables: cause, event name, victims, related places, proper names and related feature types.
- 4 = The disambiguated placename matches the attributes and the other way around. It must include attributes from Event and Disaster tables: cause, event name plus any other.
- 3 = The disambiguated placename matches the attributes and the other way around. It must include attributes from Event or Disaster tables: cause or event name plus any other.
- 2 = The disambiguated placename does not match the attributes and the other way around, but includes attributes from one of both Event or Disaster tables plus other.
- 1 = The disambiguated placename does not match the attributes and the other way around, but includes attributes only from one of both Event or Disaster tables.

Most of the marking constraints contain three aspects: number and type of attributes; quality of matching between attributes and placenames; and, Event and Disaster table completeness. The definition of the categories is thought to aids users in case they want to re-disambiguate attributes for places.

The evaluation was carried out manually taking into account the following aspects per category: First, we counted the number of text blocks; then, we identified the

times each processor handled the blocks; after, we count the frequency of the placenames at different administrative hierarchies (Table 9).

Table 9: Qualitative news ranking

Frequencies	Categories					
	1	**2**	**3**	**4**	**5**	**Total**
Blocks	13	5	27	20	5	**70**
Freeling|Geo-NER	5|8	3|2	20|7	16|4	1|4	**45|25**
State|Mun|Loc	5|0|8	2|1|2	14|11|2	11|8|1	4|1|0	**36|21|13**

Table 9 summarizes the findings per block text in each category; including how many times they appear (*Blocks* label). Also the times whether blocks were analyzed by dependency grammar or by Geo-NER (*Freeling*|*Geo-NER* label). Also summarizes how many times entities from the different administrative levels are found (*State*|*Municipalities*|*Localities* label).

Interpretation

If we split Table 9 between two parts, and in the first we include categories *one* and *two*, and in the second part we include categories *three* to *five*; the first comprises ~26 % of inaccurate placename-attribute matching. That means the relationship between attributes and placenames mentioned in blocks belonging to category one-two is senseless. The second group counts roughly ~74% of blocks corresponding to categories three-fifth; they may be considered “more accurate” because their attributes listed match with the “correct” placenames.

Second row in Table 9 notice that dependency analysis (Freeling) was performed more times than Geo-NER. The use of Freeling was raising proportionally in categories *from one* to *four*, 38, 60, 74, 80% respectively; however, last proportion suddenly decreased in the *fifth* category (20%). Conversely, the use of Geo-NER decreased from category one to fourth, and suddenly increased in the fifth category reaching 80% of use.

Regarding the administrative levels of placenames, half of the output places resulted in state entities. The remaining proportion is shared by municipalities (30%) and localities (20%). Municipalities are majority concentrated in categories *three* and *four*. Most of the localities entities were concentrated on category *one*.

Discussion

Here we discuss some of the findings from the test we performed.

In the general analysis (Table 8) we detected that some of the irrelevant news mentioned nameplaces but their context was senseless for our interest. For example

they mention political facts or events related to sports news. Likewise, other stories reported concepts related to natural hazards and impacts of disasters, but the places lay outside the NGF boundaries.

Both processors seems to have performed well. Most stories with highest ranks were treated by Geo-NER, and most of stories in categories three and four by Freeling. Despite this, both processors still faced some drawbacks. For example, in category *one*, 7 out of 8 localities tried by Geo-NER were highly complex. The processor could not solve the non-geo ambiguity even by using the placenames codes as a filter. The complexity remains because the attributes refer to a feature types not to placenames. The only placename described matches exactly with the name of an organization (Figure 16). On the other hand, the fact that Freeling uses dependency grammar to assign attributes to their corresponding places, it also would not be able to disambiguate the same case, because it uses Geo-NER for distilling placenames.

Heavy rains in northern and central Mexico keep in alert civil protection authorities, as many dams are at maximum capacity and the *National Water Commission* (Conagua) has had to increase the venting

---

Placename = National Water Commission
Attribute = Heavy rains (Event)
Organization = National Water Commission
Organization = Placename

Figure 16: Complex story

Regarding the granularity of the output placenames, conversely to what we expected, most of them were *state* entities, followed by *municipalities*. The identification and extraction of such entities, also relies on the complexity in the news stories. Many news items refer to feature types but they mention only the placenames of higher hierarchies of feature types they belong; for example in Figure 17.

The Government Secretariat (GS) declared *emergency* in *10 municipalities* of *Durango* affected by *heavy rain*, where two *cyclists* were *killed* and several *bridges* were toppled by the force of water.

Placename = Durango ***<-state***
Attributes = Emergency (Event), heavy rains (Cause), cyclists killed, bridges toppled (Victims), municipalities (NGF entity) ***<-municipality***

Figure 17: Complex story II

Other stories containing event-desaster attributes do not refer to administrative placenames references to fuzzy or related places (i.e., “heavy rains in the north and

center of the country...").

The main strength of mixing Freeling and Geo-NER is to attach the extracted attributes to the right place (if many places appear in a news). However, we can not ensure completely the semantic accuracy of the output records because of the complexity mentioned before. With an unmeasurable uncertainty our contribution is to highlight whether a place has, or is vulnerable to events related to risk conditions, and in some cases detect whether a disaster has, is, or will happen. Also we can not ensure that by using our algorithm later recalls achieve the ~74% of success we got, because it depends upon the news contents. However the results of our approach may help users to populate databases faster in the way that they will have high relevant data of an specific domain.

5 Conclusions and future work

The problem of lack of spatial data can be solved by means of semi-automatic extraction of information from the Internet. As verified in the second chapter, different contributions, and serious progress has been made in the field of information retrieval and extraction. We apply some knowledge of previous work and include our own methods to solve the shortage of spatial data; focusing on the case of disasters and natural hazards.

To this end, we proposed a methodology that extends the design of a system architecture, the data model of the problem, the algorithms for data processing, as well as the evaluation of the outputs.

As for the system architecture, we designed one that is open, simple and can be implement in other use cases. We use the flow as proposed and all the components,from seeking the feeds, accessing the catalog, and then store the results in the database, so that they are used by the client.

For its part, the data model was devised for the thematic domain described in the use case. And it can be taken as a reference for the generation of other thematic models. One of its important aspects is the inclusion of NGF as a main reference for placenames. Similarly, the thematic package can be customized, additional layers can be feasibly adapted depending on the required aims.

Regarding the information extraction from web feeds, we attempted two approaches. One of them is the processor that uses named entity recognitions patterns directly to identify and categorize meaningful features in news stories. Our main contribution in this part, is the way that the processor handles the text, mainly because of the constraints to retrieve attributes and placenames. The use of the codes

of the NGF gazetteer strongly helped us achieve largely geo and non-geo disambiguation. Also at this step, we were able to get information from news to fill the database. Likewise, in this step we could check if news needed to be analyzed further.

The mixture of our custom processor along with a language analysis library is the cornerstone in this part of our work. That is because the processor is capable of handling complex text blocks under strong grammatical analysis. That allows it to assign the thematic attributes to the appropriate places. We believe that this is the first time this heuristics approach of combining grammatical tools and gazetteer codes to disambiguate placename, aiming to extract spatial information contained in news and storing it in a database, has been applied to the Spanish language.

The results obtained under the qualitative analysis carried out, show that our contribution can easily identify relevant news items on a given topic. Similarly, the results show that the quality of the records is largely compete to the news content. In that sense, although we may extract information that may be relevant to the identification and study of disasters, we can not ensure under a standard measurement the precision of the outputs. Such uncertainty occurs because semantic information in text stories is still not parametrized.

Most of the get results are at state and municipal levels. Along with the semantic problem, the scale of the spatial database must be revised by a human agent to ensure its complete reliability. Our findings jointly help to quickly identify events associated with disasters, allowing to user focus in space for more detailed attribute disambiguation that our processor could not achieve.

Future work.

Future work remains to implement the entire circuit that is based on the system and also verify their performance.

Another inclusion would be a method to track and keep up to date changes in the NGF gazetter. In the same line, increasing the resolution of the dictionary entries, including lower granularity levels such as neighborhoods and cadastre entities could be the subject of future work. Beside placenames, the addition of assets prone to to natural hazards may enhance the categories identification. This can be done by incorporating feature type gazetteers (roads, rivers) from several sources (i.e., Geonames, Geospatial Intelligence Agency, etc.). In this case of linking geodata from multiple sources, we would like to explore the possibilities of online analytical processing (OLAP).

Following the idea of [Alfonseca 2007], we would like to improve the system so that from an ontology and not from unstructured list of text used by NER, the system be

able to find instances of the ontology concepts in the input texts. The implementation can be done using the approach recommended by [Janowicz and Keßler 2008].

The records stored in the database need to be documented, we propose the incorporation of a component to automatically deduct and assign metadata. Besides documenting the TDB, also we aim to review the legal issues that allow content to be redistributed. If we can use the content without restriction, then we propose to set an open license to the TDB (i.e.,[39]). After the legal review, other enhancements include to make available TDB via a Geobrowser, as well as to incorporate a Volunteered Geographic Information section to promote public participation.

A Appendix: Placename disambiguation using gazetteer codes

Here we offer the steps taken by the Geo-NER algortihm to do perform geo and non-geo desambiguation.

Disambiguation algorithm using codes from the national geostatistic framework NGF :

1. **Identification of state entities.** The Geo-NER starts reading each text block, then
   (a) compares the names of the text against the names of states catalog
   (b) if no placename found ends
   (c) if found placename then
   (d) store them and their codes, continue on the next point
2. **Identification of municipalities**. The Go-NER iterates again over the text of the press release looking for the words municipality or county,
   (a) if doest not find the any of that words then ends and returns only the list of state placenames from the previous points
   (b) if it finds the word then
   (c) reads the names of the text and compare them with the placenames in the catalog of municipalities but only within the range of codes belonging to states entities (of step 1d)
   (d) if not found matching words in the previous step ends and
   (e) only returns the list of state entities
   (f) if found municipal toponyms then
   (g) stores them including their keys and continue
3. **Identification of localities.** The processor follows running over the news text and
   (a) reads the names of the text and compare them with place names in the list of localities, again, only within the range of the keys to the municipalities identified in step (2g)
   (b) if not find matching localities names then ends, and list only municipalities identified in step (2g), if found, then

(c) store placenames of localities with their keys

4. **Output**. This is the result after identifying placenames of states, municipalities and localities as well as attributes. The output provide a list containing:

   (a) placenames identified at any administrative level

   (b) possible attributes to fill in the tables of the data model.

5. **Filtering the output:** At this point is determined whether the output list is turned into the database or news still need to be processed with the dependency analyzer. Text title of news are excluded. Before make the decision, the processor reads the output list and verifies:

   (a) the existence of at least one attribute of event name (eveName) or cause; and a placename of any administrative level from points 1, 2 and 3.

   (b) if it found at least one entity from each of those two attributes, the iteration ends and the current text block is excluded from any post-processing, otherwise continues.

   (c) ***Disambiguation of proper names*** (non-geo disambiguation). This step seeks to improve disambiguation between proper names and placenames. To do this,

      i. the processor reads the names identified in steps 1, 2 or 3.

      ii. these names are compared against a rule, which states: “if a name of a feature type (river, street, etc.,) is preceded by a placename then is not a toponym” (i.e. Colorado River is not the same as State of Colorado in U.S.A.).

      iii. if there are place names that match the rule then

      iv. exclude names from the list

      v. otherwise continues and

   (d) using the entity-keys, the processor verify the administrative levels of the place names listed,

   (e) identifies the most detailed or lowest level in the administrative hierarchy (state> municipality> locality), then

   (f) compares that the output list ONLY has a placename in the same administrative level, if it meets this step, the list becomes tuples to populate the database (see Example 1 and point (6)), if found more than one placename in the same administrative division then

   (g) the text block must be processed with the analysis of dependencies (see Example 2).

6. **Turning the output to a Database**. If the output list met the requirements for be stored in the database (point (5f)), then, the tuples are generated from the list and turned to fill in the database. Figure (18) shows how a output list would seem like:

——————————-Placenames——————————-
found state: Michoacán de Ocampo [16]
——————————-Attributes——————————
['victim', u'muertas'] *['dead']*
['victim', u'desaparecidas'] *['missing']*
['victim', u'damnificadas'] *['affected']*
['eveName', u'inundaciones'] *['floods']*
['relFT', u'ríos'] *['rivers']*
['aroundPlace', u'en la zona oriente de Michoac\xe1n'] *['eastern part of Michoacan']*
['relFT', u'zona'] *['zone']*
['Proper', u'Michoac\xe1n'] *['Michoacan']*
['Cause', u'lluvias'] *['rain']*
——————————-common feed elements ——————————
title->Michoacán: suman 16 muertos y más de 20 mil damnificados/ *Michoacán: total 16 dead and over 20 thousand victims*
description-> Dieciséis personas muertas, –entre ellas siete niños–, más de 30 desaparecidas y 20 mil damnificadas han dejado hasta ahora las inundaciones provocadas por el desbordamiento de tres ríos en la zona oriente de Michoacán, luego de las torrenciales lluvias/
Sixteen people dead, among them seven children, over 30 missing and 20 thousand victims have so far failed to flooding by the overflow of three rivers in the eastern part of Michoacan, after the torrential rains.
pubdate->Sat, 06 Feb 2010 10:20:57 GMT
link->http://www.[...]

Figure 18: Tuples sample output

The output list in Figure (18) is parsed to create the database. Table (10) shows filled record in the database.

Table 10: Database filled from Geo-NER output

Output list	**Tuples**	**Field: Table Equivalency**
found estado: Michoacán de Ocampo [16]	estado/state	ngfFT: Place
	Michoacán de Ocampo	ngfName: Place
	16	ngfCode:Place
[set from the NGF]	the corresponding coordinates	lat:Place
[set from the NGF]	the corresponding coordinates	lon:Place
ngfCode+pubdate	'1606022010'	eventID:Event
pubdate->	'06/02/2010'	date:Event
['eveName', u'inundaciones']	'inundaciones'/ *'floods'*	eveName:Event
title->Michoacán: suman 16 muertos y más de 20 mil damnificados	'Michoacán: suman 16 muertos y más de 20 mil damnificados'/*Michoacán: total 16 dead and over 20 thousand victims*	title:Event
description-> Dieciséis personas muertas...	'Dieciséis personas muertas...'/*'Sixteen people dead...'*	desc:Event
['aroundLugar', u'en la zona oriente de Michoac\xe1n']	'en la zona oriente de Michoac\xe1n'	relPlace:Event
['relFT', u'ríos'] ['relFT', u'zona']	'ríos'/*'rivers'* 'zona'/*'zone'*	relFT:Event
['Proper', u'Michoac\xe1n']	'Michoacán'/*'Michoacan'*	proName:Event
link-> http://www.[...]	'http://www.[...]'	Source:Event
['Cause', u'lluvias']	'lluvias'/'rain'	cause:Disaster
'Impact'	not found	impact:Disaster
'Damage'	not found	damage:Disaster
['victim', u'muertas'] ['victim', u'desaparecidas'] ['victim', u'damnificadas']	'muertas'/ *'dead',* 'desaparecidas'/*"missing,* 'damnificadas'/*'affected'*	victim:Disaster
'aid'	not found	aid:Disaster

B References

[Alfonseca 2007] Alfonseca, E., F. Verdejo (ed.), Reconocimiento de Entidades, Resolución de Correferencia y Extracción de Relaciones. Acceso y viabilidad de la información multilingüe en la red: el rol de la semántica, 2007

[Ayers and Watt 2005] Ayers, D., Watt, A., Beginning RSS and Atom Programming. Wiley Publishing, Inc.10475 Crosspoint Boulevard Indianapolis, IN 46256. USA. 2005.

[Ahern et al. 2007] Ahern, Shane and Naaman, Mor and Nair, Rahul and Yang, Jeannie Hui-I: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections, JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM, 1–10, 2007.

[Amitay et al 2004] Amitay, Einat, Har'El, Nadav, Sivan, Ron, and Soffer, Aya: Web-a-where: geotagging Web content, SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 273–280, 2004.

[Badia et al. 2007] Badia, Antonio, Ravishankar, Jothi, and Muezzinoglu, Tulay: Text Extraction of Spatial and Temporal Information, ISI, IEEE, 381, 2007

[Bitrán 2000] Bitrán, D., Características e impacto socioeconómico de los principales desastres ocurridos en la República Mexicana en el periodo 1980–99. Sistema Nacional de Protección Civil Centro Nacional de Prevención de Desastres. Serie Impacto Socioeconómico, No.1, 107 p. México, 2000.

[Blaikie et al. 1994] Blaikie P, Cannon T, Davis I, Wisner B. At Risk: Natural Hazards, People's Vulnerability, and Disasters. London, UK: Routledge, 1994.

[Borges et al. 2003] Borges, Karla A. V., Laender, Alberto H. F., Medeiros, Claudia B., and Silva, Altigran S. Da: The Web as a data source for spatial databases, In Anais do V Brazilian Symposium on Geoinformatics, Campos do Jordão, 2003.

[CENAPRED 2009] CENAPRED Características e impacto socioeconómico de los principales desastres ocurridos en la República Mexicana en el año 2008. Sistema Nacional de Protección Civil Centro Nacional de Prevención de Desastres. Serie Impacto Socioeconómico, No.10, 363p. México, 2009.

[Cardoso et al. 2007] Cardoso, Nuno, Cruz, David, Chaves, Marcirio Silveira, and Silva, Mário J.: Using Geographic Signatures as Query and Document Scopes in Geographic IR, CLEF, volume 5152, Springer, 802–810, Eds: Peters, Carol, Jijkoun, Valentin, Mandl, Thomas, Müller, Henning, Oard, Douglas W., Peñas, Anselmo, Petras, Vivien, and Santos, Diana, 2007

[Carrera et al. 2008] Carrera, J., Castellón I., Lloberes, M., Padró L., and Nevena, T.: Dependency Grammars in FreeLing, Procesamiento del Lenguaje Natural, 21–8, September 2008.

[Christopher et al. 2008] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[Egenhofer et al. 1998] Egenhofer, A. R. Shariff M., Egenhofer, M., Mark, D., Rashid, A., and Shariff, A. R.: Natural-Language Spatial Relations Between Linear and Areal Objects: The Topology and Metric of English-Language Terms, 1998.

[Egenhofer et al. 1991] Egenhofer, Max J., and Franzosa, Robert D.: Point-set topological spatial relations, International Journal of Geographical Information Systems, volume 5, 161–174, 1991

[Egenhofer and Mark 1995] Egenhofer, Max J., and Mark, David M.: Naive Geography, Austria, Cosit Semmering. Springer-Verlag, 1995.

[Farrar and Lerud 1982] Farrar, R. K., and Lerud, J. V. Using geographical coordinates to search bibliographical geoscience databases. In Online ‘82 Conference Proceedings (pp. 256-262). Weston, CT: Online. 1982.

[Gibson and Erle 2006] Gibson, R. & Erle, S. Google maps hacks, Sebastopol, CA: O’Reilly. 2006.

[Hill 200] Hill, Linda L.: Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints, In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000, Springer, 280–290, 2000

[Hu and Ge 2007] Hu, You-Heng, and Ge, Linlin: A Supervised Machine Learning Approach to Toponym Disambiguation, The Geospatial Web, 117–28, 2007.

[IADB 2005] Inter-AmericanDevelopment Bank. Indicators of Disaster Risk and Risk Management Program for Latin America and The Caribbean. Summary report. World Conference on Disaster Reduction. Kobe, Hyogo, Japan, 2005.

[Janowicz and Keßler 2008] Janowicz, K., and Keßler, C.,: The Role of Ontology in Improving Gazetteer Interaction, International Journal of Geographical Information Science (IJGIS), 2008.

[Jones et al. 2002] Jones C. B., Purves, R., Ruas, A., Sester, M., Kreveld, M. Van, Kreveld, M. Van, Weibel, R.: Spatial Information Retrieval and Geographical Ontologies - An Overview of the SPIRIT Project, In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 387–388, 2002.

[Jones et al. 2008] Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H.: Modelling vague places with knowledge from the Web. International Journal of Geographical Information Science (IJGIS), 22(10), volume 22, Taylor & Francis, Inc., 1045–1065, 2008.

[Keßler et al. 2009] Keßler C. , Maué P., Heuer J. T., and Bartoschek T. Bottom-Up Gazetteers: Learning from the Implicit Semantics of Geotags. In Krzysztof Janowicz, Martin Raubal, and Sergei Levashkin: Third International Conference on GeoSpatial Semantics (GeoS 2009). December 3–4 2009, Mexico City. Springer Lecture Notes in Computer Science 5892: 83–102, 2009.

[Kuhn 2003] Kuhn, Werner: Semantic Reference Systems, International Journal of Geographical Information Science 17(5), volume 17, 405–409, 2003.

[Leidner 2007] Leidner, Jochen L.: Toponym resolution in text: annotation, evaluation and applications of spatial grounding, SIGIR Forum 41(2), volume 41, 124–126, 2007.

[Leidner et al. 2003] Leidner, Jochen L., Sinclair, Gail, and Webber, Bonnie: Grounding Spatial Named Entities for Information Extraction and Question Answering, Proceedings of the Workshop on the Analysis of Geographic References held at the Joint Conference for Human Language Technology and the Annual Meeting of the Noth American Chapter of the Association for Computational Linguistics 2003 (HLT/NAACL'03), 31–38, May 2003.

[Longley et al. 2001] Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. Geographic information systems and science. Chichester: John Wiley & Sons, Ltd. 2001.

[Manov et al. 2003] Manov, Mr Dimitar, Kiryakov, Mr Atanas, Popov, Mr Borislav, Bontcheva, Dr Kalina, Maynard, Dr Diana, and Cunningham, Dr Hamish: Experiments with geographic knowledge for informationextraction, 2003.

[Masser 2007] Masser, I., ed. Building European Spatial Data Infrastructures. Redlands, CA: ESRI Press. 2007.

[Margaret et al. 2006] Margaret Arnold , Robert S. Chen , Uwe Deichmann , Maxx Dilley , Arthur L. Lerner-Lam , Natural Disaster Hotspots Case Studies Details. The World Bank Hazard Management Unit. Washington, D.C. 2006.

[Nebert 2004] Nebert, D. D. Developing Spatial Data Infrastructures: The SDI Cookbook, Version 2.0. 2004.

[SLINERC 2002] Patrick, Jon, Whitelaw, Casey, and Munro, Robert: SLINERC: the Sydney Language-Independent Named Entity Recogniser and Classifier, COLING-02: proceedings of the 6th conference on Natural language learning, Association for Computational Linguistics, 1–4, 2002.

[Popescu et al. 2009] Popescu, Adrian, and Grefenstette, Gregory: Deducing trip related information from flickr, WWW '09: Proceedings of the 18th international conference on World wide Web, ACM, 1183–1184, 2009.

[Sallaberry et al. 2007] Sallaberry, Christian, Gaio, Mauro, Lesbegueries, Julien, and Loustau, Pierre: A Semantic Approach for Geospatial Information Extraction from Unstructured Documents, The Geospatial Web, 93–104, 2007.

[Schmid 1994] Schmid, Helmut: Probabilistic Part-of-Speech Tagging Using Decision Trees, International Conferenceon New Methods in Language Processing, 1994.

[Singhal 2001] Singhal, Amit. "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43. 2001. Avaliable at : http://singhal.info/ieee2001.pdf.

[Smith 2002] Smith, David A.: Detecting and Browsing Events in Unstructured text, SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 73–80, 2002.

[Smith and Crane 2001] Smith, David A., and Crane, Gregory: Disambiguating Geographic Names in a Historical Digital Library, ECDL '01: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries, Springer-Verlag, 127–136, 2001.

[Smith and Mann 2003] Smith, David A., and Mann, Gideon S.: Bootstrapping toponym classifiers, Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references, Association for Computational Linguistics, 45–49, 2003.

[Sundheim 1995] Sundheim, Beth M.: Overview of results of the MUC-6 evaluation, MUC6 '95: Proceedings of the 6th conference on Message understanding, Association for Computational Linguistics, 13–31, 1995.

[Stokes 2008] Stokes, Nicola, Li, Yi, Moffat, Alistair, and Rong, Jiawen: An empirical study of the effects of NLP components on Geographic IR performance, International Journal of Geographical Information Science 22(3), volume 22, 247–264, 2008.

[Turner 2008] Turner, Andrew. Where 2.0 The State of the Geospatial Web. Sebastopol, Calif.: O'Reilly Media, 2008.

[Twaroch et al. 2008] Twaroch, Florian A., Smart, Philip D., and Jones, Christopher B.: Mining the Web to detect place names, GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval, ACM, 43–44, 2008.

[UNDP 2004] UNDP, Reducing Disaster Risk a Challenge for Development, United Nations Development Programme, Bureau for Crisis Prevention and Recovery One United Nations Plaza, New York, NY 10017, USA, 2004.

[Rijsbergen 1979] van Rijsbergen C. J. . Information Retrieval. Butterworths, London, 1979. Available at: http://www.dcs.gla.ac.uk/Keith/Chapter.1/Ch.1.html

[Vestavik 2003] Vestavik, Oyvind. Geographic Information Retrieval: an Overview. 2003? Available at: http://www.idi.ntnu.no/~oyvindve/article.pdf.

[Wick and Becker 2007] Wick, Marc, and Becker, Torsten: Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization, The Geospatial Web, 105–115, 2007.

[Woodruff and Plaunt 1994] Woodruff Allison G., and Plaunt Christian GIPSY: Automated geographic indexing of text documents. Journal of the American Society for Information Science, Vol. 45. No 9, Pp, 645-655, 1994.

[Xu 2007] Xu, Jun: Formalizing natural-language spatial relations between linear objects with topological and metric properties. International Journal of Geographical Information Science 21(4), volume 21, 377–395, 2007.

[Zubizarreta et al 2008] Zubizarreta, Álvaro, de la Fuente, Pablo, Cantera, José M., Arias, Mario, Cabrero, Jorge, García, Guido, Llamas, César, and Vegas, Jesús: A georeferencing multistage method for locating geographic context in Web search, CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, ACM, 1485–1486, 2008.

C Online sources

[1] EM-DAT. Available at: http://www.emdat.be/database [Accessed November 10, 2009].

[2] Center For Hazards & Risk Research | Research: Hotspots. Available at: http://www.ldeo.columbia.edu/chrr/research/hotspots/ [Accessed October 29, 2009].

[3] Hyogo Framework - PreventionWeb.net. Available at: http://preventionweb.net/english/hyogo/national/list/?pid:23&pih:2#M [Accessed November 16, 2009].

[4] Disaster - Preparedness and Mitigation in the Americas. Available at: http://www.disaster-info.net/newsletter/107/ocha.htm [Accessed November 16, 2009].

[5] Sistema de Alerta Temprana para Centroamérica. Available at: http://www.satcaweb.org/alertatemprana/inicio/alertatemprana.aspx [Accessed November 16, 2009].calves

[6] HealthMap | Global disease alert map. Available at: http://www.healthmap.org/en [Accessed November 16, 2009].

[7] Available at: http://exploreourpla.net/explorer/ [Accessed December 8, 2009].

[8] Instituto Panamericano de Geografía e Historia - IPGH. Available at: http://www.ipgh.org/Consejo-Directivo/41-RCD/Files_41-RCD/41-RCD_Resol_01-21.pdf [Accessed November 15, 2009].

[9] OGC®. Available at: http://www.opengeospatial.org/ [Accessed October 28, 2009].

[10] Geoparsing - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Geoparsing [Accessed November 29, 2009].

[11] Geocoding - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Geocoding [Accessed November 29, 2009].

[12] Geotagging - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Geotagging [Accessed December 1, 2009].

[13] Flickr: Explore everyone's photos on a Map. Available at: http://www.flickr.com/map/ [Accessed November 11, 2009]. Extensible Markup Language (XML). Available at: http://www.w3.org/XML/ [Accessed December 1, 2009].

[14] GeoNames. Available at: http://www.geonames.org/ [Accessed November 16, 2009].

[15] GeoRSS. Available at: http://www.georss.org [Accessed November 16, 2009].

[16] Home & Abroad | Personalized Travel Planning, Dream Trips, Virtual Concierge or Browsing Attractions. Available at: http://www.homeandabroad.com/ [Accessed December 8, 2009].

[17] Flickr! Available at: http://www.flickr.com/ [Accessed December 8, 2009].

[18] Perseus Digital Library. Available at: http://www.perseus.tufts.edu [Accessed December 8, 2009].

[19] Gazetteer Britannica Online Encyclopedia. Available at: http://www.britannica.com/EBchecked/topic/227504/gazetteer [Accessed December 2, 2009]

[20] British Broadcasting Corporation (BBC) Available at: http://news.bbc.co.uk/2/hi/help/3223484.stm [Accessed January 7, 2010].

[21] GeoFeed, Geographically Aware Feeds (Atom, RDF and RSS) and Web Services. Available at: http://www.geofeed.net [Accessed November 15, 2009].

[22] Organización territorial de México - Wikipedia, la enciclopedia libre. Available at: http://es.wikipedia.org/wiki/Datos_administrativos_de_M%C3%A9xico#Estados_Mexicanos [Accessed November 28, 2009].

[23] Instituto Nacional de Estadística y Geografía (INEGI). Available at: http://www.inegi.org.mx/inegi/default.aspx [Accessed October 28, 2009].

[24] Catálogos predefinidos. Available at: http://mapserver.inegi.org.mx/mgn2k/catalogo.jsp [Accessed January 4, 2010].

[25] The Atom Syndication Format RFC4287. Available at: http://www.ietf.org/rfc/rfc4287 [Accessed December 11, 2009].

[26] The Atom Publishing Protocol-RFC5023. Available at: http://www.ietf.org/rfc/rfc4287 [Accessed December 12, 2009].

[27] XML-RPC Available at: http://www.xmlrpc.com/metaWeblogApi [Accessed December 18, 2009]

[28] Rssboard (2009), Really Simple Syndication Advisory Board: specifications, tutorials and discussion. Available at: http://www.rssboard.org/rss-specification [Accessed December 12, 2009].

[29] Rss20AndAtom10Compared Available at: http://intertwingly.net/wiki/pie/Rss20AndAtom10Compared [Accessed November 23, 2009].

[30] ACME Available at: http://www.acme.com [Accessed December 14, 2009].

[31] BBC News feed for UK in Google Maps Available at: http://dev.benedictoneill.com/bbc [Accessed January 7, 2010].

[32] La Jornada > Aviso legal. Available at: http://www.jornada.unam.mx/aviso.php [Accessed November 12, 2009].

[33] CONAPO Available at: http://www.conapo.gob.mx/index.php?option=com_content&view=article&id=78&Itemid=194 [Accessed January 18, 2009].

[34] SRTM Available at: http://www2.jpl.nasa.gov/srtm/ [Accessed January 18, 2009].

[35] Getty (Gazetter) Thesaurus of Geographic Names. Available at: http://www.getty.edu/research/conducting_research/vocabularies/tgn/ [Accessed September 18, 2009].

[36] FreeLing User Manual. Available at: http://garraf.epsevg.upc.es/freeling/doc/userman/html/ [Accessed November 10, 2009].

[37] Python: Regular expression operations Available at: http://docs.python.org/library/re.html [Accessed January 7, 2010].

[38] Desinventar Organization Available at: http://www.desinventar.org/en/projects/promoter [Accessed October 30, 2009].

[39] Open Data Commons » Licenses. Available at: http://www.opendatacommons.org/licenses/ [Accessed November 12, 2009].

[40] Message Understanding Conference. Available at: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html [Accessed December 19, 2009].

[41] Aggregator - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Aggregator [Accessed November 20, 2009].

[42] CGI: Common Gateway. Available at: Interface</http://hoohoo.ncsa.illinois.edu/cgi/intro.html [Accessed February 14, 2010].

[43] LingPipe: Competition Available at: http://alias-i.com/lingpipe/Web/competition.html [Accessed November 25, 2009].

[44] Geographic Search and Referencing Solutions - MetaCarta - At the Forefront of the GeoWeb Available at: http://metacarta.com/ [Accessed December 12, 2009].

[45] Information extraction - Wikipedia, the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Information_extraction [Accessed December 19, 2009].

[46] Language-Independent Named Entity Recognition (II). Available at: http://www.cnts.ua.ac.be/conll2003/ner [Accessed January 27, 2010].