On the use of data mining to improve the knowledge of historical earthquakes

WORK PACKAGE 2 - EARTHQUAKE PARAMETERS

AUTHORS		REVIEW		APPROVAL	
Name	Date	Name	Date	Name	Date
Emmanuelle Nayman *Meryl Bothua* *Jessie Mayor* *Natacha Testut*	*2020/09/25*	Thierry Camelbeeck RMW Musson	2020/10/13 12 Oct 2020	Emmanuel VIALLET Public access ✓ SIGMA-2 restricted	2020/12/21

***DISSEMINATION**: This document must not be distributed to any person, institution or company other than members of SIGMA-2 steering and scientific committees, except under written formal permission of SIGMA-2 steering committee.*

Document history

DATE	VERSION	COMMENTS
2019/10/08	*1*	*First draft submitted*
2020/09/25	*2*	*Revised version*

Executive summary

In recent years growth of digital data has been increasing, and the World Wide Web (www) is the most heterogeneous and dynamic repository available. This work proposes a framework based on data mining techniques to improve knowledge of historical earthquakes by finding new records on literary heritage available on the web.

Data mining technology helps to extract useful information from various databases. Data mining on text has been designated at various times as statistical text processing, knowledge discovery in text, intelligent text analysis, or natural language processing, depending on the application and the methodology that is used ([12]).

The method presented here focuses on a specified available corpus of documents: Gallica©, the digital library of the Bibliothèque nationale de France (BnF, [9]).

The main part of this work focuses on designing methods and algorithms in order to effectively process this avalanche of text. To guarantee the success of such a process and define a precision strategy, three key steps are highlighted here:

- **Exploiting existing databases**: Exploiting the macroseismic database SISFRANCE (BRGM-EDF-IRSN; [7]), where about 10 000 bibliographic references have been collected to describe 6 000 earthquakes (463-2007), seismological ontology is defined and used as dedicated dictionary to extract relevant information from the Gallica© collection of documents.
- **Semantic enrichment of databases and knowledge enrichment**: The collection of documents are text data, which can be defined as unstructured information. Gallica© documents are semantically annotated with seismological ontology (dedicated dictionary) but also with named entities (regular expressions such as cities, dates or numbers) which constitute the knowledge base. Gallica© documents are thus turned into structured information.
- **Using advanced techniques of data mining:** the use of a similarity process dramatically helps to find relevant text through the background 'noise'.

The proposed methodology aiming at finding new sources to improve past earthquake knowledge is not destined to replace historian expertise on documents themselves. Expert assessment, by analyzing and interpreting sources, and putting them into the historical context is crucial. This methodology presented here just aims at facilitating source findings and delivering new sources in their hands.

Table of Contents

1. Motivation

- **Why improving past EQ knowledge?**

Metropolitan France belongs to the western European intraplate domain and behaves as a rigid block characterized by low internal deformation rates ([1], [2]). In such a context, the instrumental seismicity is characterized as low to moderate. However strong earthquakes occurred in the past (see **Figure 1**). This seismogenic behavior of geological structures induces a very long return period for the biggest events. In mainland France, one strongly destructive seism and four seisms creating severe damages occur within a one-thousand-year period. As example, we can list the Lambesc earthquake at the beginning of the 19th century or the Bâle earthquake in the 14th century with magnitudes between 6.0 and 7.0.

Seismometer networks in Metropolitan France able to record strong motion are only a few decades old, the first deployment of seismometers began in 1962. The estimation of hazard must cope with the small amount of instrumental data available which is not representative of seismic activity in mainland France.

To overcome this limitation, it is essential to resort to historical seismicity which allows to cover a larger time window and to include longer return period events when performing robust seismic hazard assessment studies.

When studying historical seismicity and thereby macroseismic data, there are only observations from a limited number of locations available for many earthquakes, particularly for those that occurred over a century ago. To refine past earthquake knowledge, new observations need to be considered.

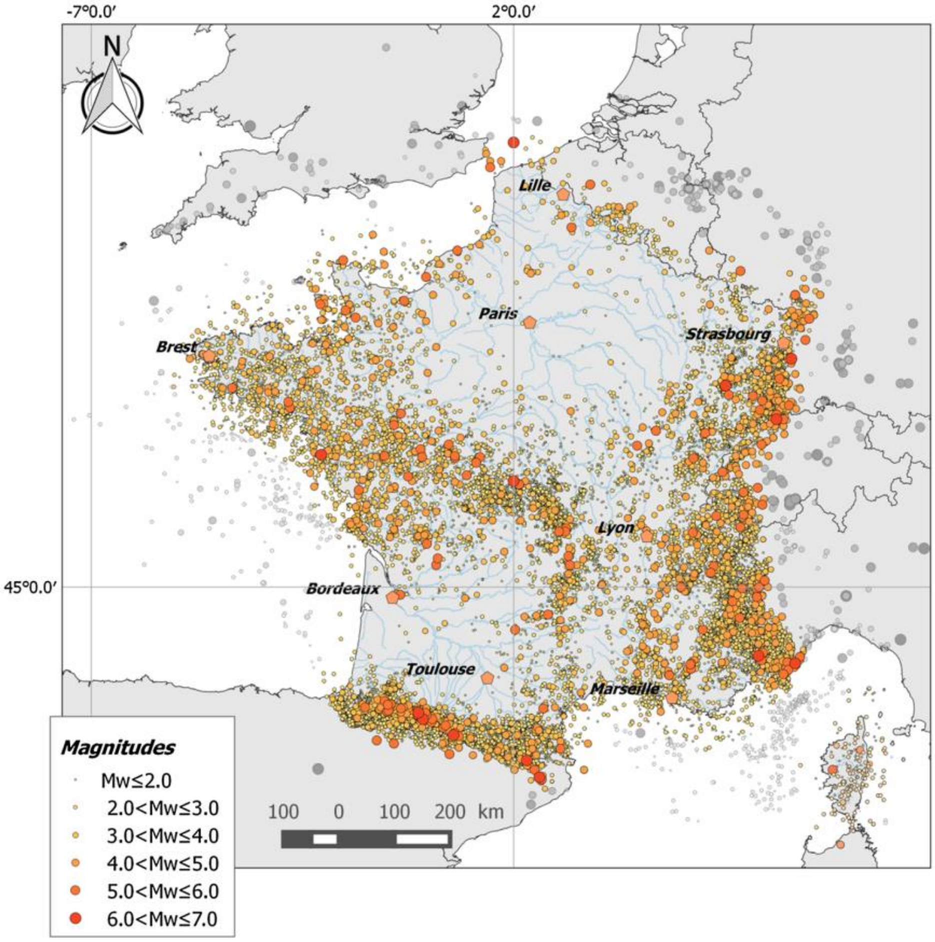


Figure 1 - The French seismic CATalogue (FCAT-17). Size and color of circles are defined according to magnitude values. [3]

- **Why using data mining techniques?**

In recent years growth of digital data has been increasing, knowledge discovery and data mining have attracted great attention and therefore created the need to turn such data into useful information and knowledge.
Moreover, many universities, government agencies, and historical associations provide digital libraries of primary sources on the Internet.

The World Wide Web (www) is thus becoming the most heterogeneous and dynamic repository available.
It is that this volume of text available on the web is an invaluable source of information and knowledge. As a result, there is a real need to design methods and algorithms in order to effectively process this avalanche of text in a wide variety of applications and to transform unstructured data into structured data, to find relevant text (testimonies on earthquakes felt in mainland France) through the background ‘noise’ (all other documents).

2. Strategy & key steps

2.1. Definition of the project

Outlines

This new methodology aiming at finding new sources to improve past earthquake knowledge is not destined to replace historian expertise on documents themselves. Expert assessment, by analyzing and interpreting sources, and putting them in their historical context is crucial.
Furthermore, it can’t replace the need for historians to visit archives to look for documents.
The methodology presented in this paper just aims at facilitating source findings and delivering new sources in their hands.

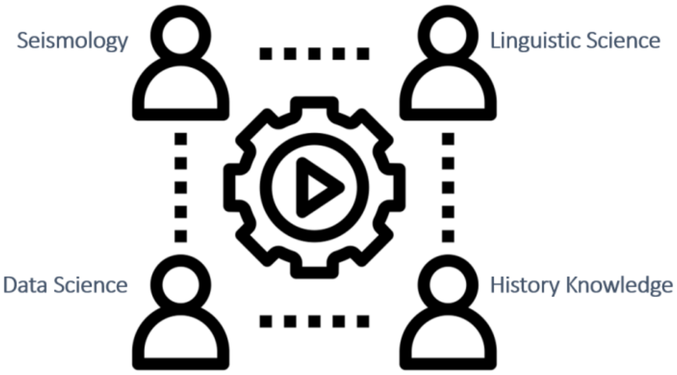
A multidisciplinary project

Given the challenge of this project, different disciplines and professions need to be combined to guarantee its success. Actors with specific knowledge such as seismological, linguistic, computer science and historian culture are then engaged in working together as equal stakeholders in addressing a common challenge: improving past earthquakes knowledge using **data mining techniques.**

eDF Engineering dept.
R&D dept.

QWAM Innovative start-up

Université Paul Valéry Montpellier University



- **EDF and its different departments** (engineering and research & development department): Seismologists and data scientists specialized in text mining methods.

- **University of Paul Valery, Montpellier**: Data analyst and historian knowledge.

- **QWAM, an innovative start-up specialized in semantics and artificial intelligence**

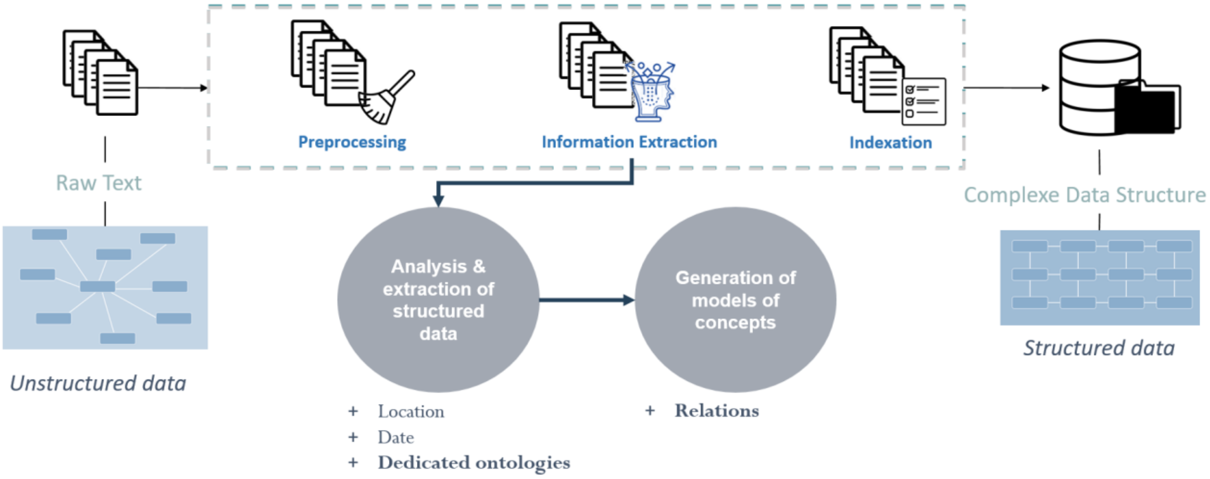
  Since its creation in 2007, QWAM works for companies and organizations by helping them with a better use of information assets and feeds, whether it is about external information (web sites, web news, blogs, etc.) or internal information (reports and studies, contracts, HR, CRM, R&D and so forth).

  QWAM's mission is to supply innovative solutions devoted to the management and analysis of unstructured information (textual big data), whether it resides within the organization or outside on the internet (https://en.qwamci.com/).

2.2. Strategy

Build and investigate a system allowing the exploitation of massive collections of documents.

Collections of documents are text data which can be defined as unstructured information. They are one of the simplest forms of data. They are easily processed and perceived by humans, but it is significantly harder for machines to understand them. The main part of this work focuses on designing methods and algorithms in order to effectively process this avalanche of texts. After collecting documents, the first challenge is to transform these raw texts into a database with predefined fields on which we can realize requests to find relevant documents. This is the text mining phase, to transform unstructured data into structured one. A preprocessing is first applied to clean the texts at best. Then, relevant information is extracted such as location, date and dedicated ontologies. This allows to build models of concepts such as relations between categories of ontologies.



Dedicated ontologies will be created using existing databases and will be used in fine to find relevant relations between concepts which fit descriptions of seisms we can observe in testimonies at the best.

2.3. Key steps

2.3.1. Exploiting existing database

Exploiting the macroseismic database SISFRANCE (BRGM-EDF-IRSN; [7]), where about 10 000 bibliographic references have been collected to describe 6 000 earthquakes (463-2007), the seismological ontology is defined and used as dedicated dictionary to extract relevant information from the Gallica© collection of documents.

See section 4, Learning from existing database

2.3.2. Semantic enrichment of database and knowledge enrichment: ontologies and dictionaries

A collection of documents are text data, which can be defined as unstructured information. Raw data (harvested documents) are semantically annotated with seismological ontology (dedicated dictionary) but also with named entities (regular expression such as cities, dates or numbers) which constitute the knowledge base. Harvested documents are thus turned into structured information, allowing information extraction.

See section 5.2, Information /concept Extraction, indexation

2.3.3. Using advanced techniques of data mining

By answering the question: *how 'close' texts in existing databases and collected texts are in meaning or surface closeness?* similarity methods are implemented and dramatically help to find relevant text through the background ‘noise’. This technique reveals hidden connections.

The “bag-of-words” assumption, presented in this work, is one of the most popular vectorization models for the similarity process. It considers a piece of text (or a document) as a set of words. In this assumption, the sequence of words is ignored, only their existence matters.

See section 5.3, Similarity - Bag of Words

3. Choice of database to explore through data mining techniques

3.1. Reasons motivating the choice of Gallica©

Given that the web provides a great quantity of documents, this current work focuses on a specified available corpus of documents: Gallica©, the digital library of the Bibliothèque nationale de France (BnF), mainly for three reasons:

Reason 1: Online collection of documents are very important
Almost 4 million documents are available on the Gallica© website. It seems to be a representative sample of documents to validate or not this method of finding new archive documents on past earthquakes using data mining techniques.

Reason 2: This collection of documents includes relevant documents on past earthquakes.
Documents including records on past earthquakes were found manually while surfing on this web site.

Reason 3: Post processing on online collection
Gallica© benefits from the progress made in Optical Character Recognition (OCR) technology. A growing number of documents were consequently digitized both in image and text modes. As a result, searches within the digital library search system became more efficient and comprehensive.

3.2. Gallica© Overview

Gallica© [https://gallica.bnf.fr] is the digital library of the Bibliothèque nationale de France (BnF, [8], [9]): a digital encyclopedia containing printed materials (books, journals, newspapers, printed music, and other documents), graphic materials (engravings, maps, photographs, and others), and sound recordings.

Gallica© makes it possible to find sources that are rare, unusual, out-of-print, or difficult, if not impossible, to access. These materials are royalty-free and available free of charge if used strictly for private purpose. This digital library includes more than 70,000 volumes of digitized texts, 80,000 still images, and 30 hours of sound recordings.

3.3. Harvesting collection of documents

Given the large number of online documents on the Gallica© website, a massive collect method is required.
The automatization of the Gallica© online collection harvesting is executed by a program or automated script which browses the entire website (**Figure 2**). In a methodical, automated manner this process will search for the relevant information using algorithms that narrow down the search by finding out the closest and relevant information. This program is called web crawler.
For more information on crawl techniques, please refer to [16] and [17].

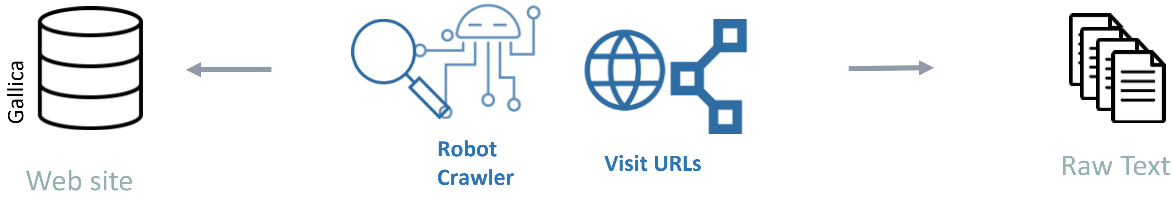


Figure 2 - Crawl of Gallica© documents

Specific requirements for Gallica© documents harvesting

- ✓ Required data: written documents themselves and linked metadata information;
- ✓ Non oriented selection: we want to retrieve the entire collection of documents from Gallica© without selection criteria (no use of key words in Gallica© advanced search engine): no a priori selection;

- ✓ Preprocessed documents: we want to retrieve documents which benefit from OCR processing. In other words, documents which are available in PDF and TEXT format to ease the text mining process.

Finally, more than 3.8 million documents need to be collected from the Gallica© website.

Tools

The harvesting of Gallica© documents is performed by QWAM's solution called Ask'nRead© module. It focuses on **extracting online news and information** and filtering it through **multilevel categorizations**. Ask'n'Read© comes with a **powerful search engine** to retrieve relevant data. It is available in SaaS mode (Software as a Service).

All documents and metafiles are downloaded in a data warehouse system. All information are text data and therefore in raw format. It corresponds to unstructured information, which is one of the simplest forms of data that can be generated in most of the cases.

The next challenge is to discover knowledge from all these data and to structure information to be able to query on this knowledge.

With this integral crawl of Gallica© we make sure not to miss any potential records on past earthquakes, but it requires to find a methodology to dramatically filter all this corpus of documents.

It will be necessary:

- ✓ To eliminate noisy documents without any link with the objectives of this study, which only drowns relevant documents in the corpus,
- ✓ To identify relevant documents containing records on past earthquakes.

4. Learning from existing database

Constitute a seismological ontology which will be used in text mining extraction process chain as filter

4.1. SisFrance database

4.1.1. General Presentation

SisFrance ([7]) is the current name for the macroseismic database originally named SIRENE, which was created in 1975 by the consortium between BRGM (Bureau de Recherches Géologiques et Minières), EDF (Électricité de France) and IRSN (Institut de Radioprotection et de Sûreté Nucléaire).

BRGM handles the management, the updating and the interpretation of the macroseismic information contained in SisFrance.

The SisFrance macroseismic database contains about 100,000 macroseismic observations (IDP, Intensity Data Point) associated with about 6000 earthquakes (AD463-2007). All intensities in the database have been evaluated with the Medvedev–Sponheuer–Karnik 1964 intensity scale ([5]).

Epicentral location is determined and provided, together with the epicentral intensity value if possible (see [6] for epicentral location and intensity assessment explanations).

Earthquake characteristics, such as location and intensity but also observations are associated with quality factors that reflect confidence related to numerical value.

Epicentral location are associated with quality factor named QPOS, whose value can be:

- quality A: certain location (accurate to a few kilometers),
- quality B: fairly certain location (accurate to 10 kilometers),
- quality C: uncertain location (accurate to 20 kilometers),
- quality D: fairly uncertain location (accurate to 50 kilometers),
- quality E: arbitrary location,
- quality I: location resulting from only one observation.

Figure 3 shows the distribution of earthquakes in SisFrance database according to this quality factor, QPOS. A small proportion of these events (~14%) are defined with a certain epicentral location (accurate to 10 kilometers).

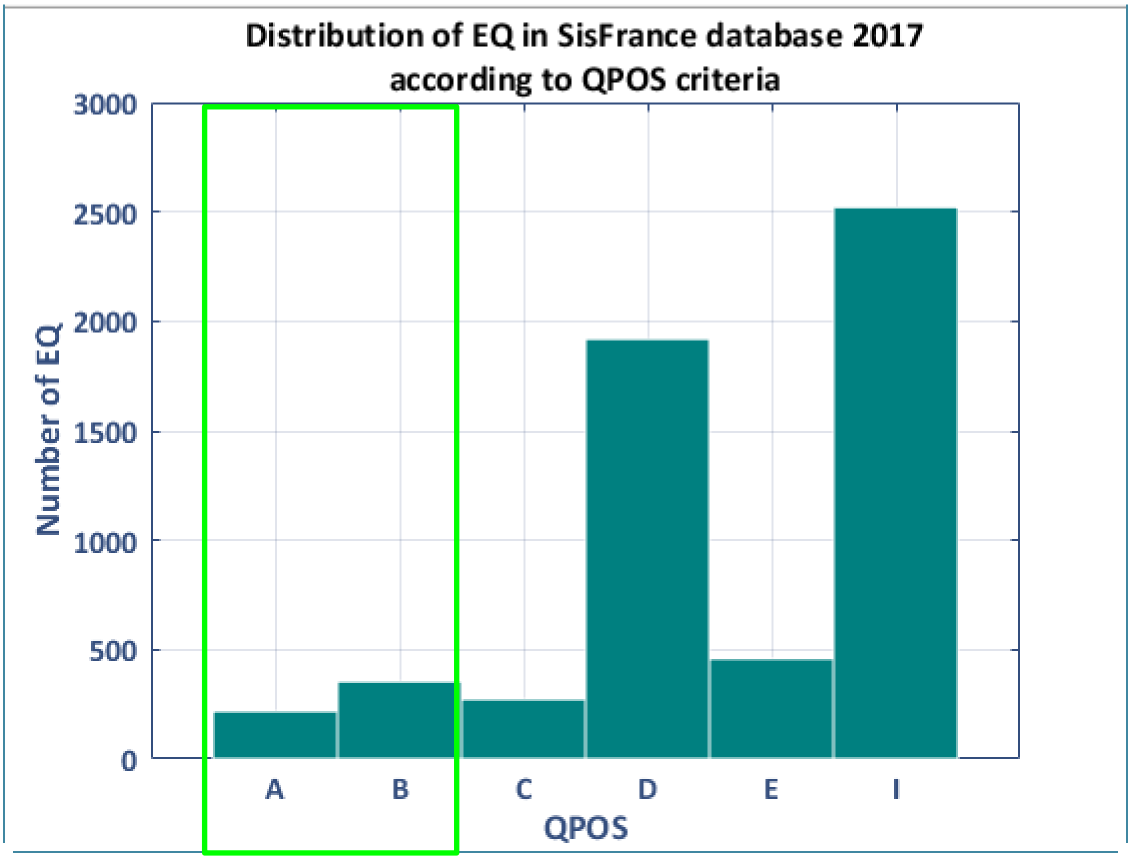


Figure 3 - Distribution of EQ in SisFrance 2017 database according to QPOS criteria

Epicentral intensity estimates are also associated with quality factors named QIE, whose value can be:

- quality A: certain intensity,
- quality B: fairly certain intensity,

- quality C: uncertain intensity,
- quality K: resulting from a calculation based on intensity attenuation,
- quality E: arbitrary intensity,
- quality I: intensity resulting from only one observation.

In addition, some observations simply state that the event was felt at that site but there is insufficient information to assign an intensity value.

Figure 4 shows the distribution of earthquakes in SisFrance database according to this quality factor, QIE. A small proportion of these events (~20%) are defined with a certain epicentral intensity (QIE ≥ B).

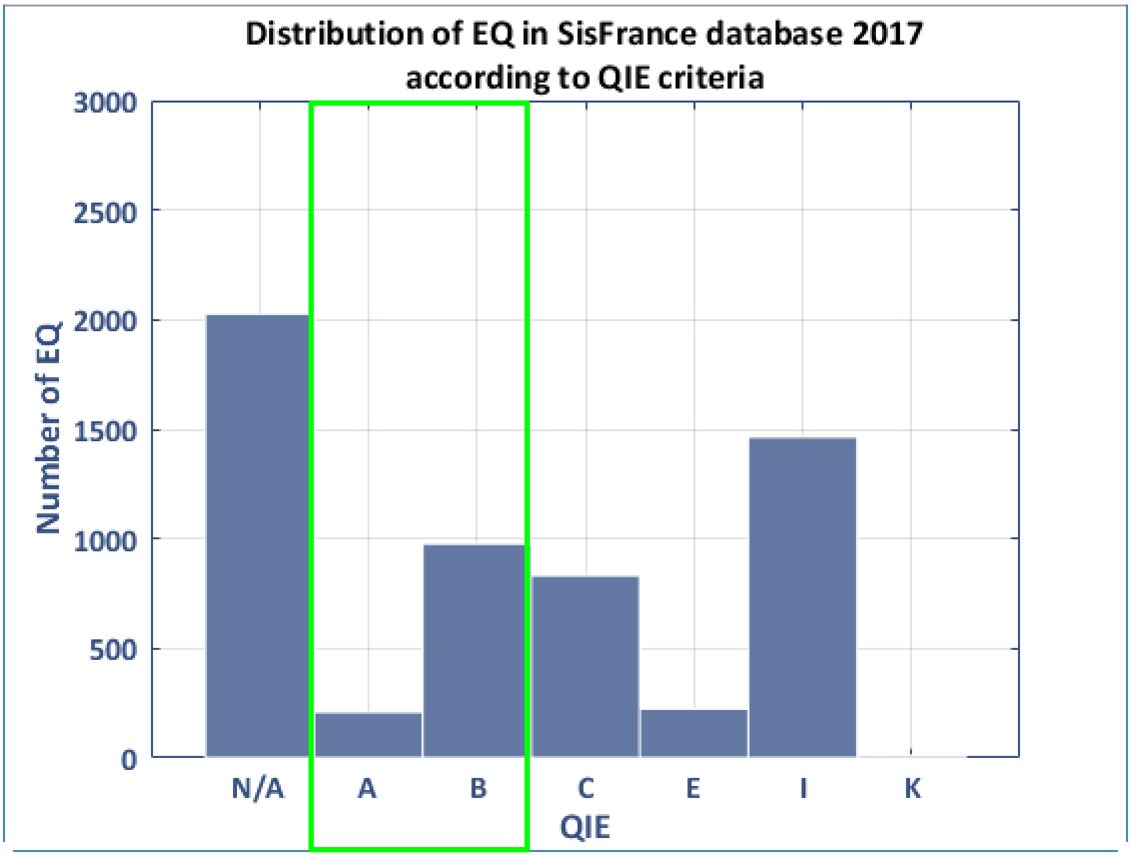


Figure 4 - Distribution of EQ in SisFrance 2017 database according to QIE criteria

These figures need to get closer to the number of IDPs describing an earthquake.

Figure 5 shows the distribution of earthquakes in SisFrance 2017 database according to the number of IDPs describing earthquakes. More than 60% of the earthquakes are not well constrained (described by less than 3 IDPs).

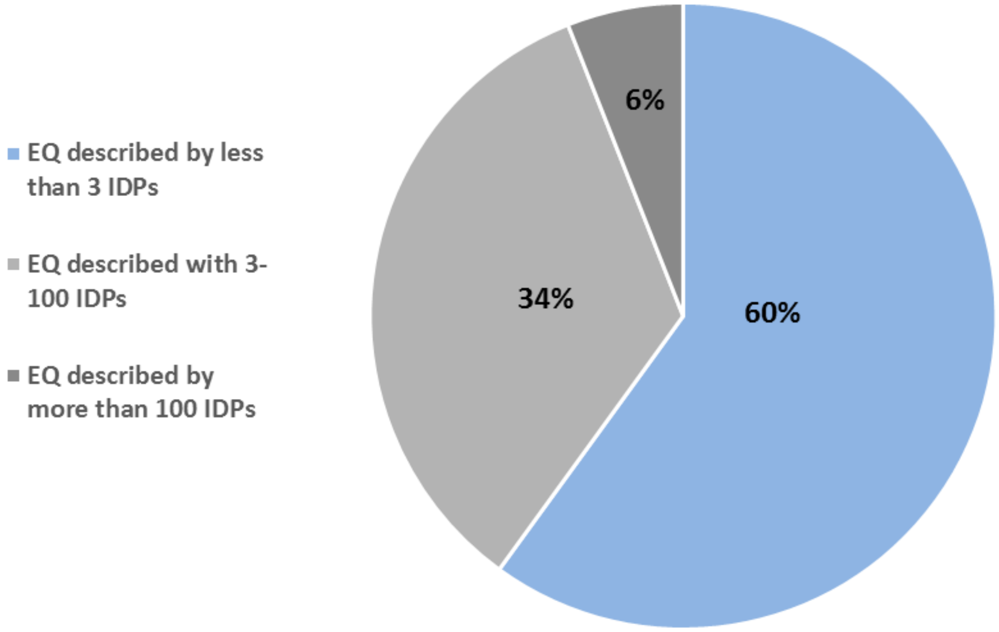


Figure 5 - Distribution of EQ in SisFrance 2017 database according to number of IDPs

4.1.2. Assessments and Needs

When studying historical seismicity, no quantitative data are available and therefore no direct access to seismological parameters such as magnitude or depth. Estimating characteristics of past earthquakes in terms of location and magnitude are a real challenge and need the study of macroseismic data.

Intensity data points (IDPs) are the only form of numerical data available for seismologists. The derivation of earthquake parameters from macroseismic (intensity) data is thus an inveterate problem.

Two criteria are essential to assess: the number and the quality of IDP, and therefore the number and quality of the records (see **Figure 6**). For many past earthquakes, especially those which occurred over a century ago, only observations at a limited number of locations are available. The exact spatial extent of the area where these earthquakes were felt will never be known.

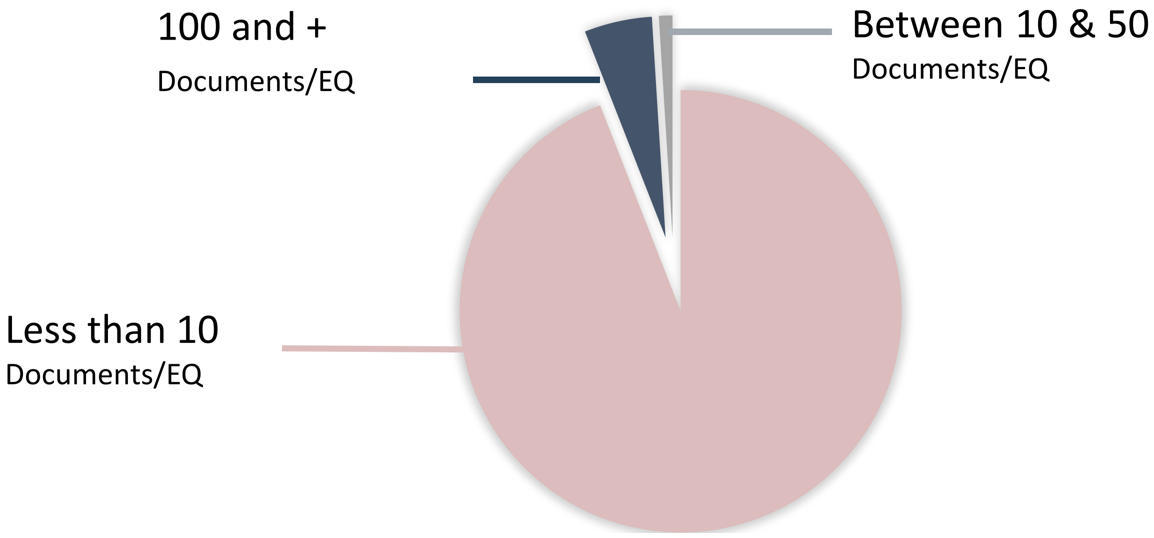


Figure 6 - Distribution of EQ in SisFrance 2017 database according to number of IDPs

All these observations lead to the same conclusions that other channels need to be involved to find new documents in order to improve past earthquake knowledge.

Database population

Some past earthquakes are better known than others and the reliability of earthquake-parameters depends on both the number and the quality of archive documents (parish registers, press clipping …) coming from different geographical locations where the earthquakes were felt. Indeed, having a maximum of information on past earthquakes is crucial to estimate robust epicentral intensity (and magnitude) and location.

When new information appears regarding earthquakes already recorded in the database, obtained by careful examination and analysis of newly identified historical documents (e.g. city records or local accounts, departmental and national archives as well as newspapers and other historical publications), it is added. In this case, the new information is compared to previously existing documents to reevaluate the characteristics of the event, sometimes leading to the inclusion, modification or suppression of IDPs.

Up to now, the most common way to find new historical sources has been the work of historians appointed to investigate on one or more earthquakes (see **Figure 7**).

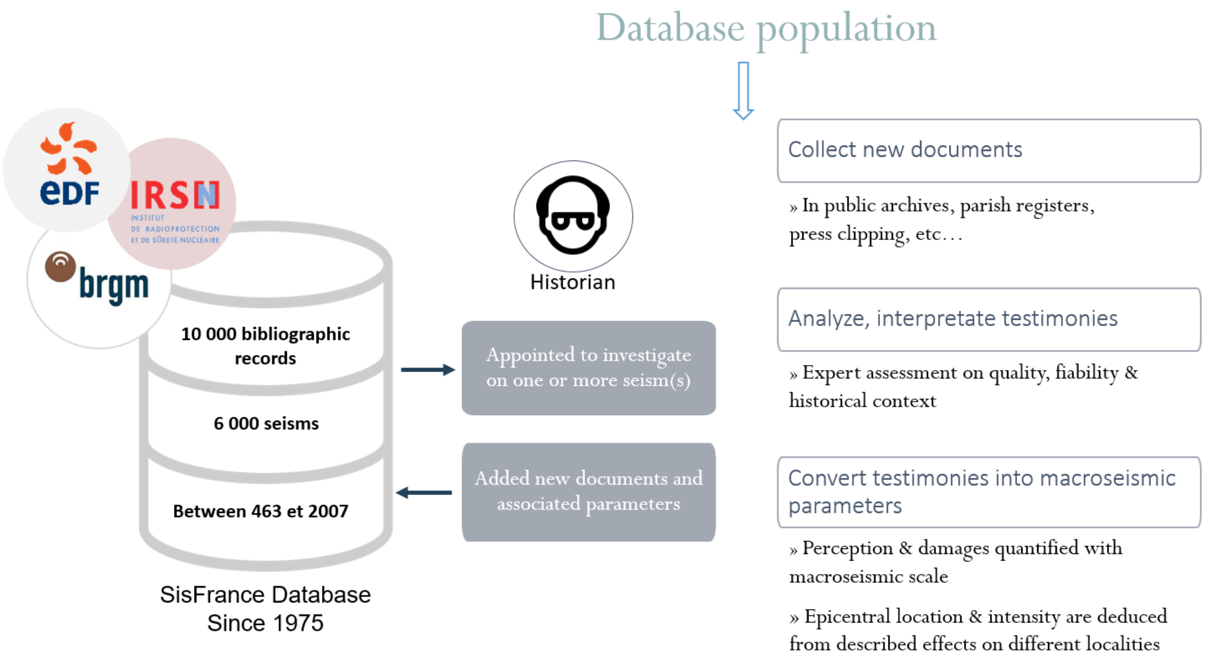

Database population
EDF
IRSN INSTITUT DE RADIOPROTECTION ET DE SÛRETÉ NUCLÉAIRE
brgm
10 000 bibliographic records
6 000 seisms
Between 463 et 2007
SisFrance Database
Since 1975
Historian
Appointed to investigate on one or more seism(s)
Added new documents and associated parameters
Collect new documents
» In public archives, parish registers, press clipping, etc…
Analyze, interpretate testimonies
» Expert assessment on quality, fiability & historical context
Convert testimonies into macroseismic parameters
» Perception & damages quantified with macroseismic scale
» Epicentral location & intensity are deduced from described effects on different localities

Figure 7 - SisFrance database and population process

4.1.3. Distribution of bibliographic records in terms of nature

All historical sources stored in the SisFrance database are characterized as primary source. They are all identified, and an ID named CHRONO is assigned to each document. In the same way, an ID named NUMEVT is assigned to each seismic event.

Primary sources are immediate, first-hand accounts of a topic, from people who had a direct connection with it.
Primary sources can include:

- Other original documents;
- Newspaper reports, by reporters who witnessed an event or who quote people who did;
- Speeches, diaries, letters and interviews - what the people involved said or wrote.

Figure 8 shows the distribution of primary sources according to the nature of documents in the SisFrance 2017 database.

On the other hand, a secondary source of information is one that was created *later* by someone who *did not* experience first-hand or participate in the events or conditions you're researching. They can cover the same topic as the primary sources, but add a layer of interpretation and analysis.
Secondary sources can include:

- Most books about a topic;
- Analysis or interpretation of data;
- Scholar or other articles about a topic, especially by people not directly involved.

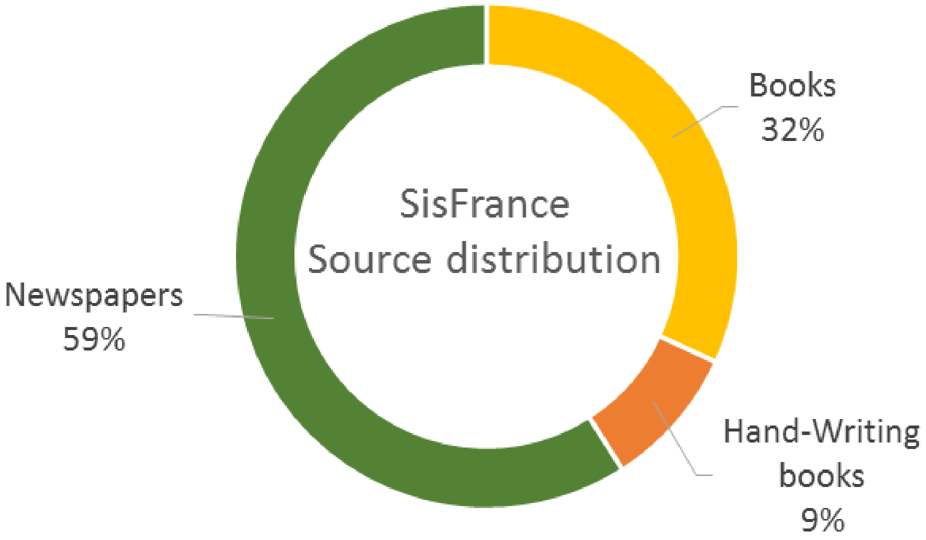


Figure 8 – SisFrance 2017: source distribution

4.2. Creation of seismological ontology

Given the large amount of harvested texts from the Gallica website, a simple search engine is useless to find relevant text through the background 'noise'. The retained strategy to collect massively relevant documents is to put a document into vector space and define its own latent space.

In order to do so, a dedicated dictionary must be created to focus on documents dealing with earthquakes felt in Metropolitan France.

Thanks to expert knowledge, six lexicons containing 206 terms were defined from SisFrance documents. They constitute a seismological ontology which will be used in the text mining extraction process chain as filters: extraction of dedicated concepts, detection of relation inter-concepts.

4.2.1. Definition of a seismological Ontology

Six concepts from earthquake vocabulary which constitutes a dedicated seismology ontology, are defined:

- **Seismic** Concept: containing all words or technical verbal phrases related to seisms,
- **Damage** Concept: containing all words or technical verbal phrases related to material or physical damages,
- **Assembly** Concept, related to building structure
- **Behavior** Concept: containing all words or technical verbal phrases related to human and animal behavior before and after earthquake occurrence and their perception,
- **Noise** Concept: containing all words or technical verbal phrases related to sound heard before and after earthquake occurrence,
- **Divine** Concept, gathering all divine allusions.

This categorization will be useful for the detection of relations between concepts and refine the filters in the text mining extraction process chain.

At the beginning, we manually created different lists given the importance of the concepts. It is a very constraining task that takes a lot of time without ensuring exhaustiveness. Furthermore, we had to deal with dirty data. In order to automatize the process and make it more efficient, we made use of the SisFrance database.

4.2.2. Using OCR to Convert Documents

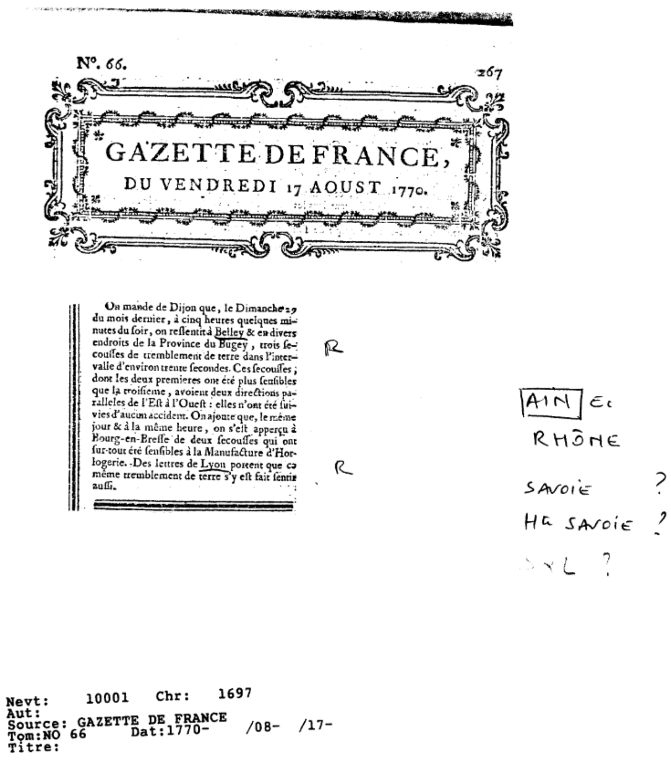
The original format of SisFrance documents is PDF. In order to extract information, we had to use an Optical Character Recognition tool (OCR) named Tesseract.

OCR is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text mainly from a scanned document or a photo of a document. It is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as text mining. For more detailed information on OCR technologies, please refer to [18].

Tesseract was in the top three OCR engines in terms of character accuracy in 1995 [19]. The initial version of Tesseract could only recognize English-language text. Tesseract v2 added six additional Western languages (French, Italian, German, Spanish, Brazilian Portuguese and Dutch). Version 3 extended language support significantly to include ideographic (Chinese & Japanese) and right-to-left (e.g. Arabic, Hebrew) languages. V3.04, released in July 2015, added an additional 39 language/script combinations, bringing the total count of support languages to over 100.

Tesseract's output will have very poor quality if the input images are not preprocessed to suit it. A lot of documents from SisFrance database were not easy to manage due to their quality but we were able to extract relevant information. An example of a document coming from SisFrance can be seen below. We can observe that Tesseract had difficulties to deal with the document's quality but relevant information had been preserved (**Figure 9**).

In other cases, original document quality can be very poor and OCR software cannot preserve anything (**Figure 10**)



N°. 66. 267

GAZETTE DE FRANCE,
DU VENDREDI 17 AOUST 1770.

On mande de Dijon que, le Dimanche 29 du mois dernier, à cinq heures quelques minutes du ſoir, on reſſentit à Belley & en divers endroits de la Province du Bugey, trois ſecouſſes de tremblement de terre dans l'intervalle d'environ trente ſecondes. Ces ſecouſſes; dont les deux premieres ont été plus ſenſibles que la troiſieme, avoient deux directions paralleles de l'Eſt à l'Oueſt : elles n'ont été ſuivies d'aucun accident. On ajoute que, le même jour & à la même heure, on s'eſt apperçu à Bourg-en-Breſſe de deux ſecouſſes qui ont ſur-tout été ſenſibles à la Manufacture d'Horlogerie. Des lettres de Lyon portent que ce même tremblement de terre s'y eſt fait ſentir auſſi.

R

R

AIN Ec

RHÔNE

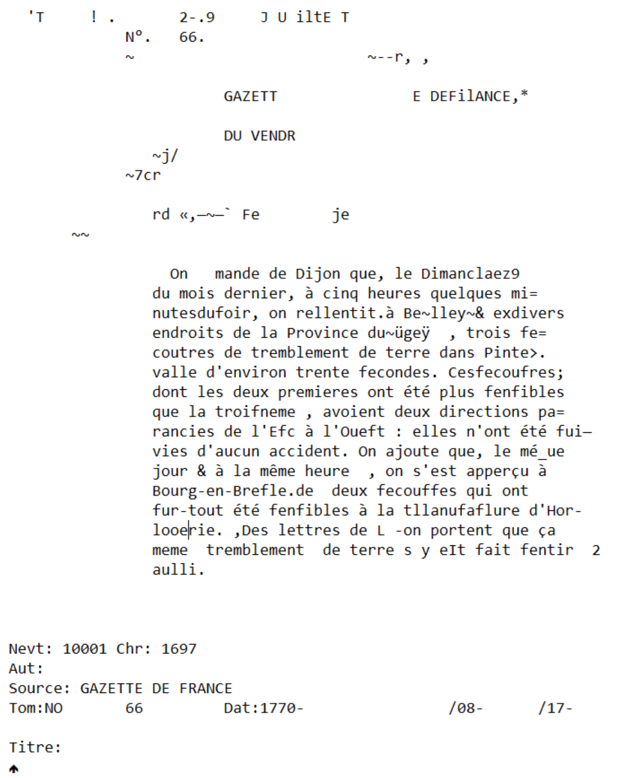
SAVOIE ?

Hte SAVOIE ?

YL ?

```
Nevt:    10001   Chr:    1697
Aut:
Source: GAZETTE  DE  FRANCE
Tom:NO  66        Dat:1770-     /08-   /17-
Titre:
```

Original document

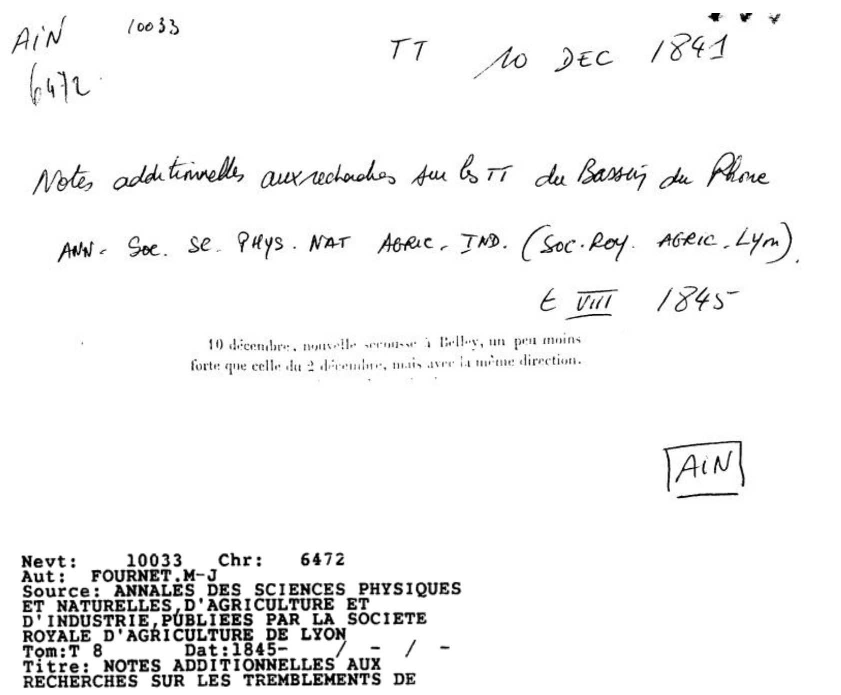


```
 'T     ! .       2-.9     J U iltE T
           N°.   66.
           ~                          ~--r, ,

                      GAZETT                E DEFilANCE,*

                      DU VENDR
              ~j/
           ~7cr

              rd «,—~—` Fe        je
      ~~

                On   mande de Dijon que, le Dimanclaez9
              du mois dernier, à cinq heures quelques mi=
              nutesdufoir, on rellentit.à Be~lley~& exdivers
              endroits de la Province du~ügeÿ  , trois fe=
              coutres de tremblement de terre dans Pinte>.
              valle d'environ trente fecondes. Cesfecoufres;
              dont les deux premieres ont été plus fenfibles
              que la troifneme , avoient deux directions pa=
              rancies de l'Efc à l'Oueft : elles n'ont été fui—
              vies d'aucun accident. On ajoute que, le mé_ue
              jour & à la même heure  , on s'est apperçu à
              Bourg-en-Brefle.de  deux fecouffes qui ont
              fur-tout été fenfibles à la tllanufaflure d'Hor-
              looerie. ,Des lettres de L -on portent que ça
              meme  tremblement  de terre s y eIt fait fentir  2
              aulli.



Nevt: 10001 Chr: 1697
Aut:
Source: GAZETTE DE FRANCE
Tom:NO       66         Dat:1770-                /08-      /17-

Titre:
```

Obtained Result

On mande de Dijon que, le Dimanche 29 du mois dernier, à cinq heures quelques minutes du soir, on ressentit à Belley et en divers endroits de la Province du Bugey, trois secousses de tremblement de terre dans l'intervalle d'environ trente secondes. Ces secousses, dont les deux premières ont été plus sensibles que la troisième, avaient deux directions parallèles de l'Est à l'Ouest : elles n'ont été suivies d'aucun accident. On ajoute que, le même jour & à la même heure, on s'est aperçu à Bourg-en-Bresse de deux secousses qui ont sur-tout été sensibles à la Manufacture d'Horlogerie. Des lettres de Lyon portent que ce même tremblement de terre s'y est fait sentir aussi.

Figure 9 - OCR output



AiN 10033
6472

TT 10 DEC 1841

Notes additinnelles aux recherches sur les TT du Bassin du Rhone

ANN. Soc. Sc. PHYS. NAT AGRIC. IND. (Soc. Roy. AGRIC. Lyon).

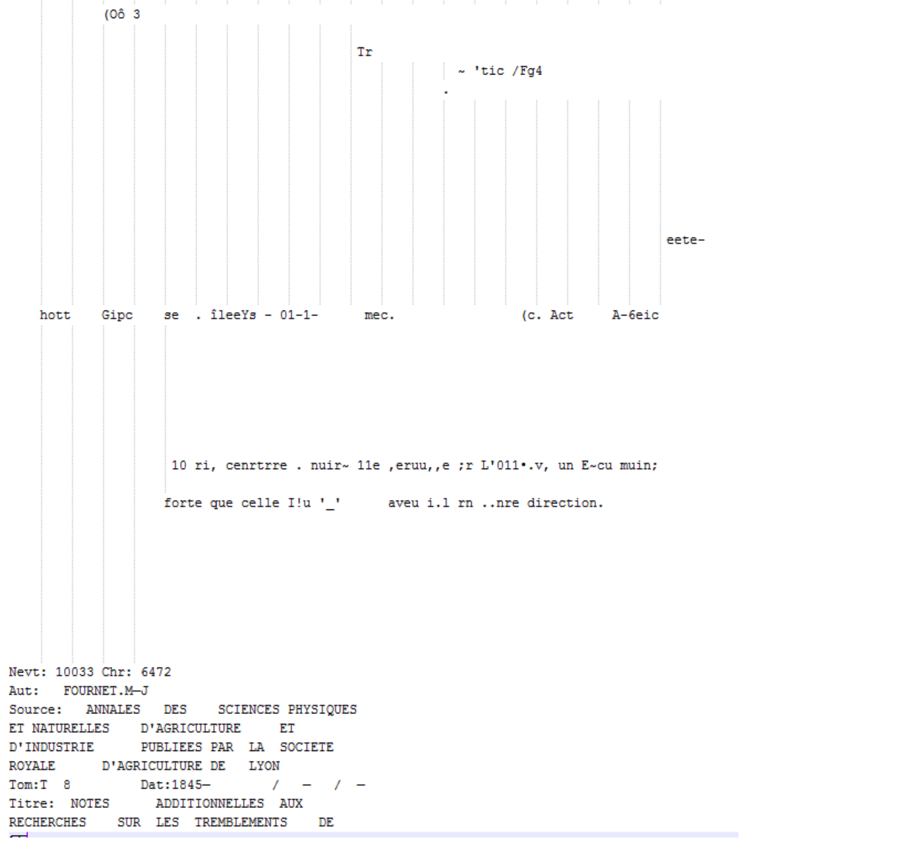
t VIII 1845

10 décembre, nouvelle secousse à Belley, un peu moins
forte que celle du 2 décembre, mais avec la même direction.

AiN

```
Nevt:    10033   Chr:    6472
Aut:  FOURNET.M-J
Source: ANNALES DES SCIENCES PHYSIQUES
ET NATURELLES,D'AGRICULTURE ET
D'INDUSTRIE,PUBLIEES PAR LA SOCIETE
ROYALE D'AGRICULTURE DE LYON
Tom:T 8          Dat:1845-     /  -  /  -
Titre: NOTES ADDITIONNELLES AUX
RECHERCHES SUR LES TREMBLEMENTS DE
```

Original document



```
        (Oô 3

                                   Tr
                                              ~ 'tic /Fg4
                                             .



                                                                  eete-


hott    Gipc    se  . îleeYs - 01-1-     mec.             (c. Act    A-6eic






        10 ri, cenrtrre . nuir~ 11e ,eruu,,e ;r L'011•.v, un E~cu muin;

        forte que celle I!u '_'      aveu i.1 rn ..nre direction.
```

```
Nevt: 10033 Chr: 6472
Aut:   FOURNET.M—J
Source:   ANNALES   DES    SCIENCES PHYSIQUES
ET NATURELLES    D'AGRICULTURE     ET
D'INDUSTRIE      PUBLIEES PAR  LA  SOCIETE
ROYALE      D'AGRICULTURE DE   LYON
Tom:T  8         Dat:1845—       /   —   /  —
Titre:  NOTES      ADDITIONNELLES  AUX
RECHERCHES    SUR  LES  TREMBLEMENTS    DE
```

Obtained Result

10 décembre, nouvelle secousse à Belley, un peu moins forte que celle du 2 décembre, mais avec la même direction.

Figure 10 - OCR output

4.2.3. Language Detection to work on French Documents

The SisFrance database was not organized by language of the documents. It was important for us to be able to organize the documents by language and to isolate French documents. In order to do so, we used an n-gram method.

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text. The items can be phonemes, syllables, letters and words. An n-gram of size 1 is referred to as a unigram, size 2 is a bigram and size 3 is a trigram.

In our case, we used a four-gram. For example if the document is “data mining development.” and n = 4, the function will return: "data" =1, “min” = 1, “ing d”=1, “eve”=1, “lopm” “ent.”.

We then developed an algorithm to compare the n-grams of a new document with the n-gram profile of each language based on sample documents considering the frequency of each n-gram in the document. For that we had a reference corpus of 6 languages: French, Old French, German, English, Spanish, and Latin. We obtained the results below (**Figure 11**).

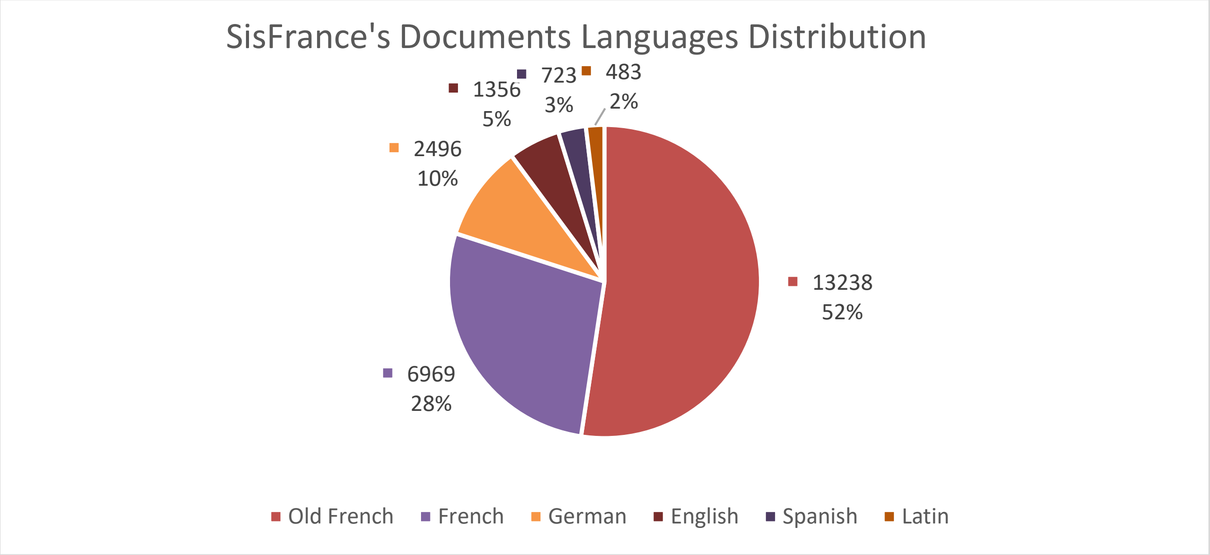


Figure 11 - SisFrance Documents Languages Distribution

4.2.4. Exploiting SisFrance documents content to enrich lexicons with word embeddings

After having developed a process dedicated to the extraction of earthquake vocabulary, and sorted the data base by languages, we decided to use the SisFrance database to complete our lexical resources. As explained previously, the documents of Gallica as those of SisFrance include many lexical irregularities due to OCR errors and lack of language standardization (a lot of synonyms); we wanted to extract these words automatically.

In order to do so, we developed a web application, CuriosiText (**Figure 12**). It is based on a neural network Word2Vec identifying similar terms used in a same context in documents.

Word2vec [21] was created and published in 2013 by a team of researchers led by Tomas Mikolov at Google; it produces word embeddings according to linguistic contexts of words. Each word of a document is converted into a vector. Word vectors are positioned in the vector space in a way that words which share common contexts in the corpus are located close to one another in the space.

Based on the method, CuriosiText suggests terms identified as similar: agglutinated words, spelling mistakes, abbreviations and various synonyms. Users can select interesting terms to add to the ontology. This ontology can next be used to extract information from documents.

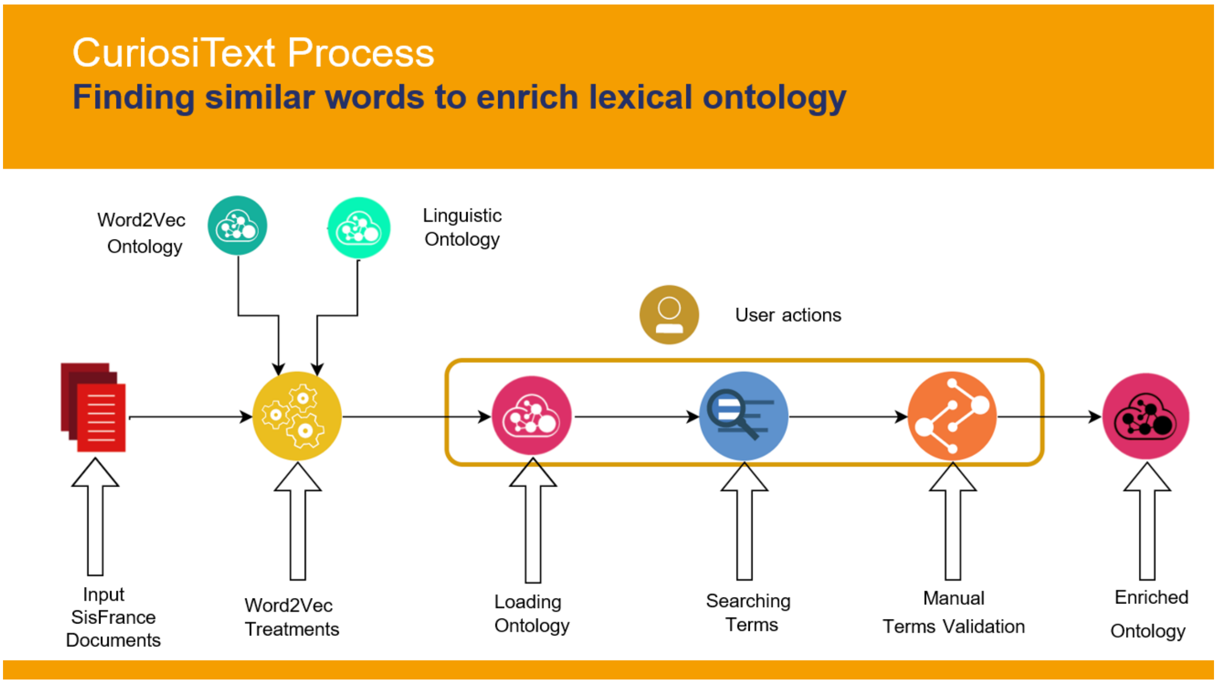
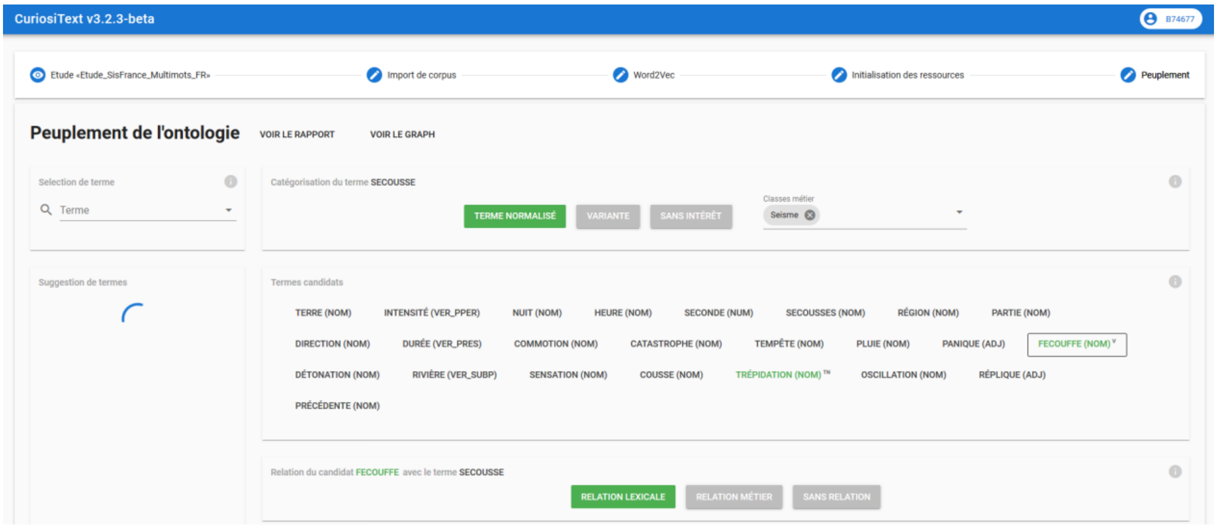


Figure 12 - CuriosiText process

CuriosiText is well adapted to any document, regardless of the language. Once the corpus has been loaded, pretreatments are applied such as the deletion of undesired characters or stopwords, morphosyntactic tagging (identifying, verb, noun, adjective etc.) and filtering by word frequency.

Then, CuriosiText calculates the similarity between word vectors thanks to the cosine method. User can load their predefined ontology containing concept (see **Figure 13**) and add relevant terms.

We illustrate below the obtained result for the word “secousse”:



CuriosiText v3.2.3-beta   B74677

Etude «Etude_SisFrance_Multimots_FR» — Import de corpus — Word2Vec — Initialisation des ressources — Peuplement

Peuplement de l'ontologie   VOIR LE RAPPORT   VOIR LE GRAPH

Selection de terme

Terme

Catégorisation du terme **SECOUSSE**

TERME NORMALISÉ   VARIANTE   SANS INTÉRÊT

Classes métier: Seisme

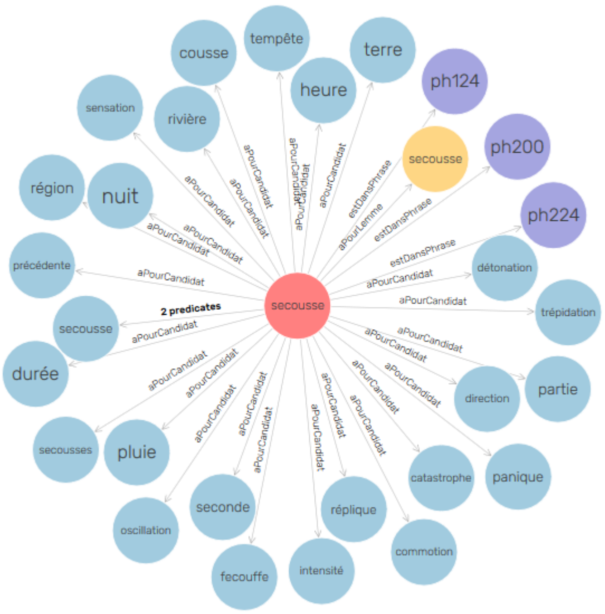
Suggestion de termes

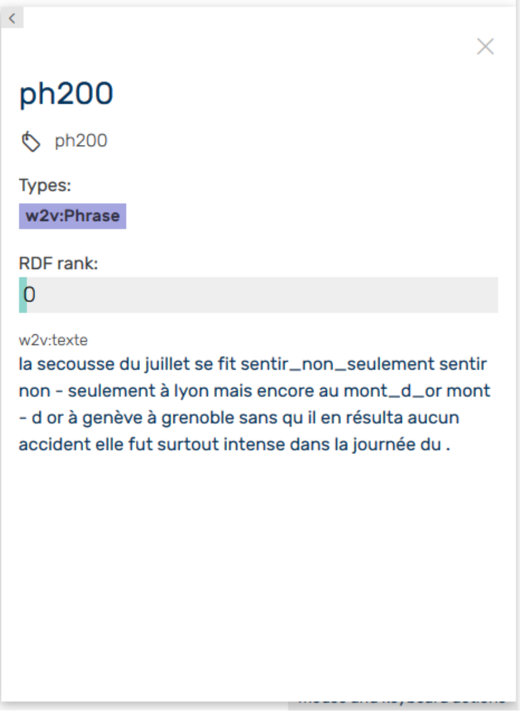
Termes candidats

TERRE (NOM)   INTENSITÉ (VER_PPER)   NUIT (NOM)   HEURE (NOM)   SECONDE (NUM)   SECOUSSES (NOM)   RÉGION (NOM)   PARTIE (NOM)
DIRECTION (NOM)   DURÉE (VER_PRES)   COMMOTION (NOM)   CATASTROPHE (NOM)   TEMPÊTE (NOM)   PLUIE (NOM)   PANIQUE (ADJ)   FECOUFFE (NOM)
DÉTONATION (NOM)   RIVIÈRE (VER_SUBP)   SENSATION (NOM)   COUSSE (NOM)   TRÉPIDATION (NOM)   OSCILLATION (NOM)   RÉPLIQUE (ADJ)
PRÉCÉDENTE (NOM)

Relation du candidat FECOUFFE avec le terme **SECOUSSE**

RELATION LEXICALE   RELATION MÉTIER   SANS RELATION





ph200

ph200

Types:

w2v:Phrase

RDF rank:

0

w2v:texte

la secousse du juillet se fit sentir_non_seulement sentir non - seulement à lyon mais encore au mont_d_or mont - d or à genève à grenoble sans qu il en résulta aucun accident elle fut surtout intense dans la journée du .

Figure 13 - Similar words obtained for “secousse” with word embeddings method

4.2.5. Ontology Enrichment

Thanks to that method, we were able to enrich predefined lexicons and to improve the accuracy of the text mining process chain. We added 109 terms to the 206 terms identified by experts.

For example, we found in some SisFrance documents the words “sacristie” (vestry) as Assembly Concept, “excavation” (hollow) as Damage Concept, “pleistoseiste” as Seismic Concept, “panique” (panic) as Behavior Concept and “canonnade” as Noise Concept.

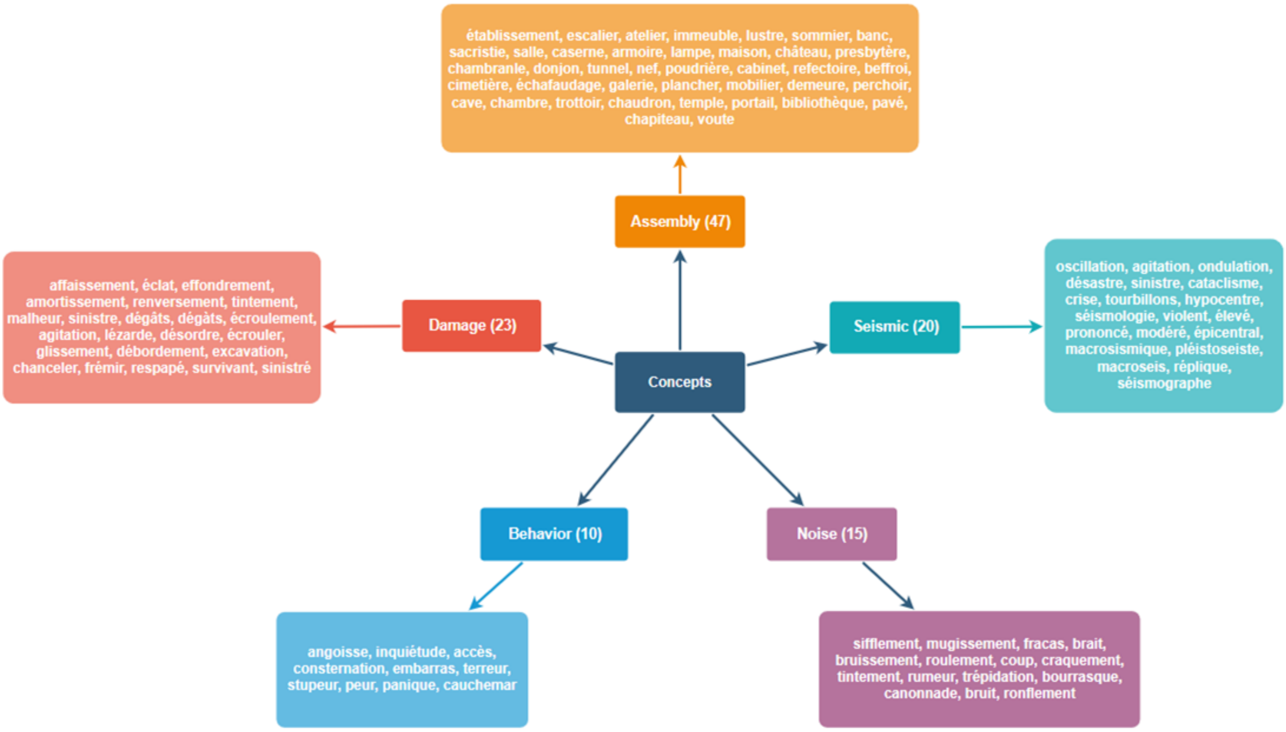

établissement, escalier, atelier, immeuble, lustre, sommier, banc, sacristie, salle, caserne, armoire, lampe, maison, château, presbytère, chambranle, donjon, tunnel, nef, poudrière, cabinet, refectoire, beffroi, cimetière, échafaudage, galerie, plancher, mobilier, demeure, perchoir, cave, chambre, trottoir, chaudron, temple, portail, bibliothèque, pavé, chapiteau, voute
Assembly (47)
affaissement, éclat, effondrement, amortissement, renversement, tintement, malheur, sinistre, dégâts, dégàts, écroulement, agitation, lézarde, désordre, écrouler, glissement, débordement, excavation, chanceler, frémir, respapé, survivant, sinistré
Damage (23)
Concepts
Seismic (20)
oscillation, agitation, ondulation, désastre, sinistre, cataclisme, crise, tourbillons, hypocentre, séismologie, violent, élevé, prononcé, modéré, épicentral, macrosismique, pléistoseiste, macroseis, réplique, séismographe
Behavior (10)
Noise (15)
angoisse, inquiétude, accès, consternation, embarras, terreur, stupeur, peur, panique, cauchemar
sifflement, mugissement, fracas, brait, bruissement, roulement, coup, craquement, tintement, rumeur, trépidation, bourrasque, canonnade, bruit, ronflement

Figure 14 – Ontology enrichment

All terms from seismological ontology are listed in Appendix 1: Seismological Ontology
The seismological ontology is thus defined and will be used as dedicated dictionary to extract relevant information from the Gallica collection of documents. This categorization will be useful for the detection of relations between concepts to refine the filters in the text mining extraction process chain.

This next step is developed in the next section.

4.3. Manual retranscription

To complete the transformation of SisFrance PDF documents into text documents, about 1000 records were manually translated.
As seen previously (**Figure 10**), OCR results can be very poor especially when facing bad print quality or hand-written records.
A choice was made to take the time to translate a set of records manually, especially older or long testimonies to get more lexicometry details for this study.
This work was not used to enrich the seismological ontology but was very useful when advanced data mining techniques were implemented such as the similarity process (see **section 5.3**).
For the most difficult cases of hand-written documents, a historian performed the translation. An example is given in the next figure (**Figure 15**).

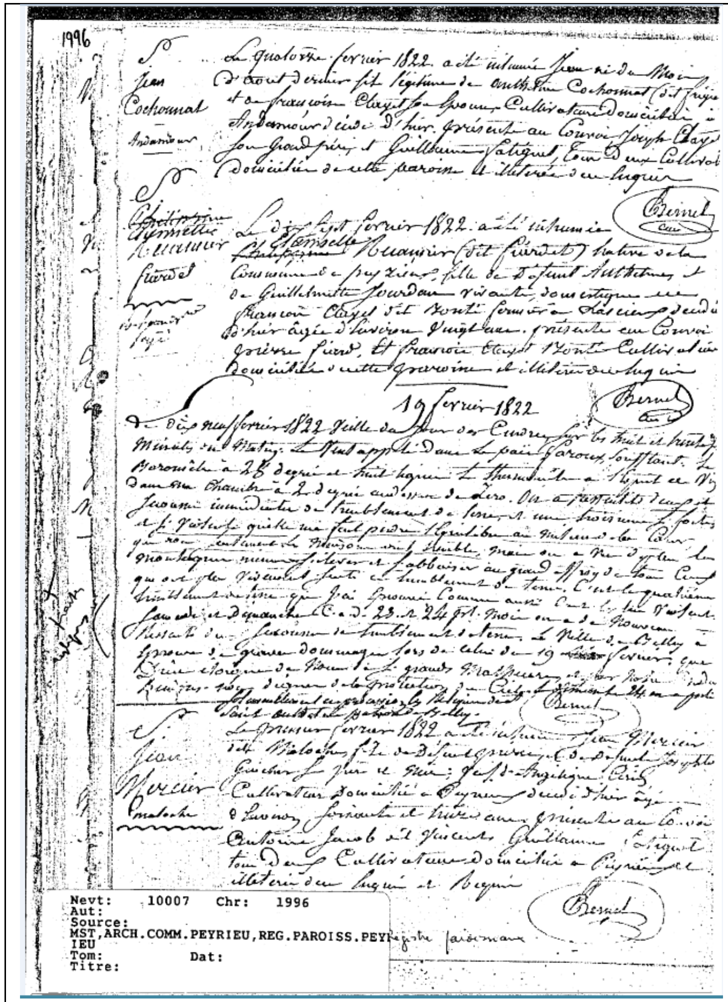
 19 février 1822 Nevt: 10007 Chr: 1996 Aut: Source: MST,ARCH.COMM.PEYRIEU,REG.PAROISS.PEYRIEU Tom: Dat: Titre:	19 février 1822 Le dix neuf février 1822 veille du jour des Cendres sur les huit et trente minutes du matin : le vent appelé dans le païs Faroux soufflant : le baromètre à 28 degrés et huit lignes le thermomètre à l'esprit de vin dans ma chambre à 2 degrés au dessus de zéro on a ressentit deux secousses immédiates de tremblement de terre, et une troisième si forte et si violente qu'elle m'a fait perdre l'équilibre au milieu de la cour que non seulement la maison ----- tremble, mais on a vu de plus les montagnes mesme se lever et s'abbaisser au grand effroy de tous ceux qui ont plus vivement senti ce tremblement de terre. C'est le quatrième tremblement de terre que j'ai éprouvé comme aussi c'est le plus violent. Samedi et dimanche c'est à dire 23 et 24 présent mois on a de nouveau ressenti deux [ou des ?] secousses de tremblement de terre. La ville de Belley a éprouvé de grands dommages lors de celui du 19 février. Que Dieu éloigne de nous si grands malheurs et par notre du ------ --- dignes de la protection du ciel le dimanche 24 on a porté solennellement en procession les reliques de Saint Anthelme patron de Belley. Besnel
Original document	Obtained Result

Figure 15 – manual retranscription for hand-writing record

5. Data Mining techniques supporting past EQ

Data mining is a process based on algorithms to analyze and extract useful information from structured data.

Text mining is the set of processes required to turn unstructured text documents or resources into valuable structured information.

Please refer to [15] for detailed information.

5.1. Preprocessing

Preprocessing is one of the key components in many text mining algorithms. For example, a traditional text categorization framework includes preprocessing, feature extraction, feature selection and classification steps. Although it is confirmed that feature extraction, feature selection and classification algorithms have significant impact on the classification process, the preprocessing stage may have noticeable influence on this success, by reducing the set of words to those that are expected to be the most relevant for the given corpus (raw text).

Preprocessing is thus essential for two main reasons:

1. To reduce indexing (or data) file size of the text documents;

2. To improve the efficiency and effectiveness of the Information Extraction (IE) system.

5.1.1. Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens. The aim of the tokenization is the exploration of the words in a sentence. The list of tokens becomes input for further processing such as parsing or text mining.

5.1.2. Lemmatization/ Stemming

It consists in two approaches to decrease the variability of words by reducing different forms of words to their basic / root form.

Lemmatization is the task that considers the morphological analysis of the words, i.e. grouping together the various inflected forms of a word so they can be analyzed as a single item.

Stemming is a crude heuristic process that chops off the ends of words without considering linguistic features of the words (for example: argue, argued, argues, arguing will become “argu”).

Lemmatization refers to the use of a vocabulary and morphological analysis of words, aiming at returning to the base or dictionary form of a word, which is known as the lemma (for example: argue, argued, argues, and arguing will become “argue”).

5.1.3. Filtering

Filtering is usually done on documents to remove some of the words. A common filtering technique is the removal of stop-words. Stop words are the words that frequently appear in the text without adding much content information (e.g. prepositions, conjunctions, etc.). Similar words appearing quite often in the text that are said to have little information to distinguish different documents and words appearing very rarely are also possibly of no significant relevance and can be removed from the documents.

Normalizing the text
The main idea is to transform various forms of the same term into a common, 'normalized' form. For example, Apple, apple, APPLE will become "apple".
By using simple rules:

- Remove all punctuation marks (dots, dashes, commas...),
- Transform all words to lower case,
- Using a dictionary, such as WordNet, to replace synonyms with a common, often more general, concept (for example: "automobile, car" will become "vehicle").

Removing terms with very small / high frequency in the given corpus
Formalized in the Zipf's rule: The frequency of a word in a given corpus is inversely proportional to its rank in the frequency table (for that corpus).
Words in the upper part of the frequency table include a significant proportion of all the words in the corpus, but are semantically almost useless (for example: the, a, an, we, do, to) (**Figure 16**).

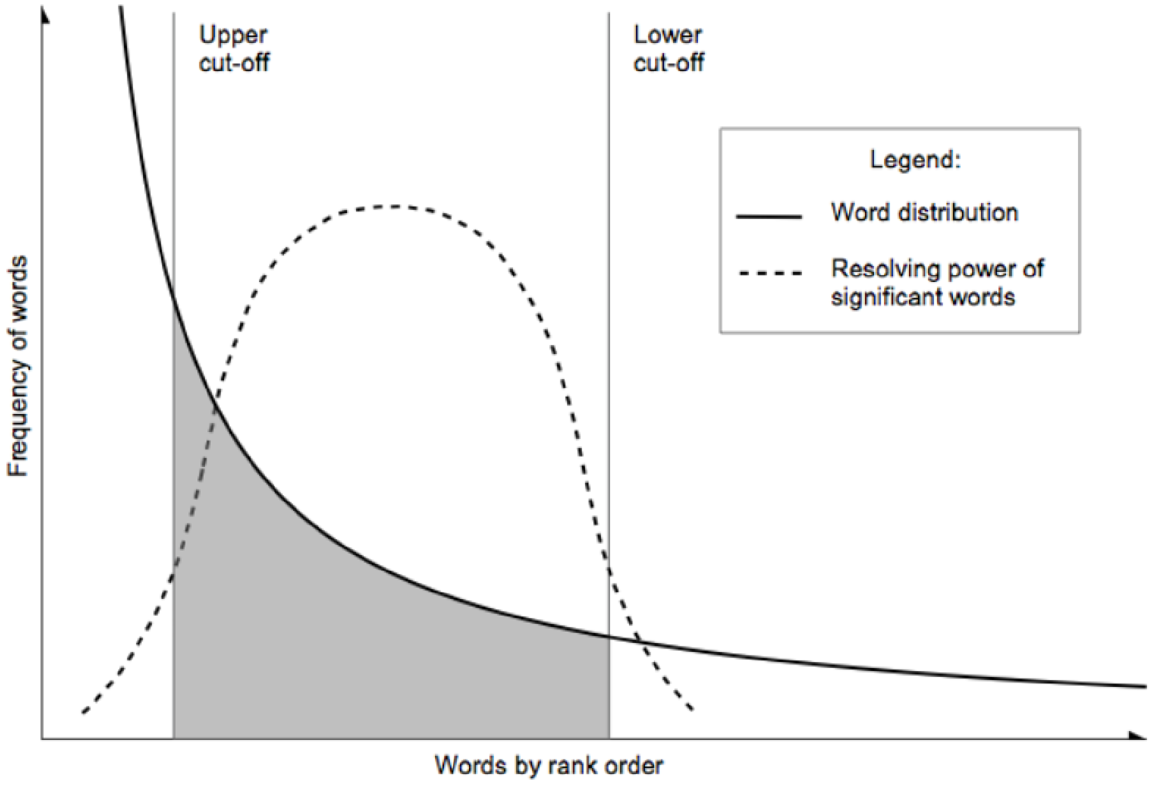


Figure 16 – Zipf's rule illustration [20]

Removing the so-called stop-words

Stop-words are those words that (on their own) do not bear any information / meaning, and are therefore irrelevant for the corpus analysis. It is estimated that they represent 20-30% of all words in any corpus. There is no unique stop-words list but frequently used lists are available at: http://www.ranks.nl/stopwords.

5.2. Information /concept Extraction, indexation

Information Extraction is the task of automatically extracting information or facts from unstructured or semi-structured documents. It usually serves as a starting point for other text mining algorithms. Information extraction includes two fundamental tasks, namely, name entity recognition and relation extraction from text who can both give us useful semantic information.

By accurately tagging all relevant concepts within a document, this method enables to rapidly identify the most relevant terms and concepts and cut through the background ‘noise’ to get to the real essence of the text.

It opens new possibilities to mine data more effectively, to derive valuable insights and to ensure not to miss anything relevant.

Standard and Discovery rules

QWAM QTA tools provide grammatical rules allowing for information extraction for named entity recognition and relation extraction using seismological entities and predefined and standard categories such as location bases (towns, countries…) or date bases.

This tool can also discover specific concepts using Word2vec techniques, globally in the same way as described in the previous **section Learning from existing database**.

Relation extraction rules

To find records on past earthquakes among this massive harvested corpus, different rules are defined to detect terms or expressions from the different branches of seismological ontology:

- ✓ Double: 2 terms or expressions from 2 seismological ontology branches available in a same sentence or section,
- ✓ Triple: 3 terms or expressions from 3 seismological ontology branches available in a same sentence or section.

These rules prove their efficiency to detect records on past earthquakes and constitute a first main reliable tool to filter documents relative to seisms among an abundance of documents. It seems logical: when people provide first-hand accounts on seism occurrence, they use several dimensions to tell their story, corresponding to the different branches of the seismological ontology.

The ontology branch called seism prevails over the other ones and is systematically coupled with other ontologies. Words such as tremor or quake (respectively “secousse” and “tremblement” in French) are the most common words used in people contribution.

Associating two or three terms from different branches of seismological ontology avoids metaphors such as “political quake” for example.

Other rules are added to render these relation extractions more efficient:

- ✓ Relations between 2 or 3 keywords are annotated with 20 words max between 2 keywords,
- ✓ Priority definition between different branches of ontology: from the highest priority to the lowest one: seism, damages, assembly, behavior, noise and divine;
- ✓ Annotation process is performed on normalized words (no accent, dash, ligature, etc…),
- ✓ Locations and dates are annotated in a relation including at least one key word from the seismological ontology branch seism;
- ✓ All relations are annotated, even if there is overlapping. The biggest one prevails on the other ones.

OCR corrections

Gallica OCR is not perfect, depending on the quality of the document itself. This induced lexicometry errors we must cope with. There is no magic formula to solve this problem and some solutions are implemented to reduce these errors. A post processing is applied on text data from OCR to increase the quality on text data on which text mining tools will be applied. As it is impossible to correct all errors, we chose to define a methodology for the 10 key words from seismological ontology for which the occurrence in the Gallica corpus is the most important. It is considered to be one of the main problems we have to solve and is subject to a great deal of attention for the next improvements.

Here are presented the 10 seismic keywords from the seismic branch of seismological ontology:

fortes secousses
légère secousse
nouvelle secousse
réplique
secousse
secousses de tremblement
seismique
tremblement
tremblement de terre
housser

The main objective is to find the regularity in terms of OCR error.
For example, if the keyword "Fortes secousses" is considered:

- Find all occurrences in text for "?ortes secousses", "F?rtes secousses", "Fo?tes secousses", etc.
- “?” can be any character,
- Same process is performed with any two characters,
- All found terms will be considered as the initial keyword and replaced for the following.

Ancient terms enrichment

As we are dealing with historical sources, the vocabulary needs to be adapted and enriched with specific terms used in historical times. The seismological ontology was enriched, especially for the seismic branch.

Two strategies are retained:

1. Translation of seismological ontology (from modern French) into old French,
2. Specific research on ancient lexical representatives of different historical time periods (Oïl language, Middle Age).

Geographical knowledge database enrichment

The geographical database used for Named Entity Recognition (NER) focuses on modern French cities. It constitutes a major problem in the frame of this work as:

- This work focuses on history, many cities were not spelled in the same way or some cities simply disappeared. This is why, we have to adapt this ontology to ancient cities from Metropolitan France to refine our search and render relation extraction more efficient;

  For this, location ontology enrichment is performed through the crawling of Wikipedia webpages

  - ✓ Enrichment with ancient cities in Metropolitan France https://fr.wikipedia.org/wiki/Listes_des_anciennes_communes_de_France

- Many major earthquakes occurred in the past and mainly outside Metropolitan France and are very well documented. These seisms don't have any interest for our study and are considered as noise. To detect these seisms and remove them, we need to enrich the location database with abroad cities to automatically discriminate them. As examples, we can cite:

  - ✓ Lisbon: 01/11/1755 - 09h40
  - ✓ Messina: 28/12/1908 - 05h20
  - ✓ Alep : 11/10/1138
  - ✓ Aleppo: 13/08/1822
  - ✓ Syria, Antioche: 13/12/115
  - ✓ Italica (Crete): 21/07/365

  For this, the location ontology enrichment is performed by crawling of the Wikipedia webpage

  - ✓ Enrichment with main cities from all over the world https://fr.wikipedia.org/wiki/Listes_des_villes_du_monde

5.3. Similarity - Bag of Words

The bag-of-words model is a way of representing text data when modeling text with machine learning algorithms: it is a way of extracting features from text for use.

Please refer to [22] for detailed information.

The main idea is to create word vectors and to score the words in each document. As the vocabulary size increases, so does the vector representation of documents.

In our context, we have to deal with two sets of documents: on the one hand the SisFrance corpus (short texts only dealing with earthquake testimonies), and on the other hand the Gallica corpus which is a very large set of documents, where earthquake records could be drowned by noise.

If a word vector is created on each Gallica document, the length of the vector might be thousands or millions of positions. But finally, each document may contain very few of the known words (seismological ontology). This result is a vector with lots of zero scores, called a sparse vector or sparse representation. Sparse vectors require more memory and computational resources when modeling and the vast number of positions or dimensions can make the modeling process very challenging for traditional algorithms.

For these reasons, word vectors are not computed on the complete Gallica document but on an abstract of this document. This abstract is automatically created from the highest density of seismological ontology terms. This process is called text summarization and is capable of extracting useful information that leaves out inessential and insignificant data. Documents are cut into paragraphs to highlight only relevant information. This extraction-based summarization is totally dependent from the quality of seismological ontology given as input.

Word vectors are thus computed between SisFrance documents and automatic abstracts from the Gallica corpus. By cosine similarity calculation (**Figure 18**), a similarity score matrix is computed varying from 0 (not similar) to 1 (similar).

The implemented similarity process chain currently under testing is synthetized in **Figure 17.**

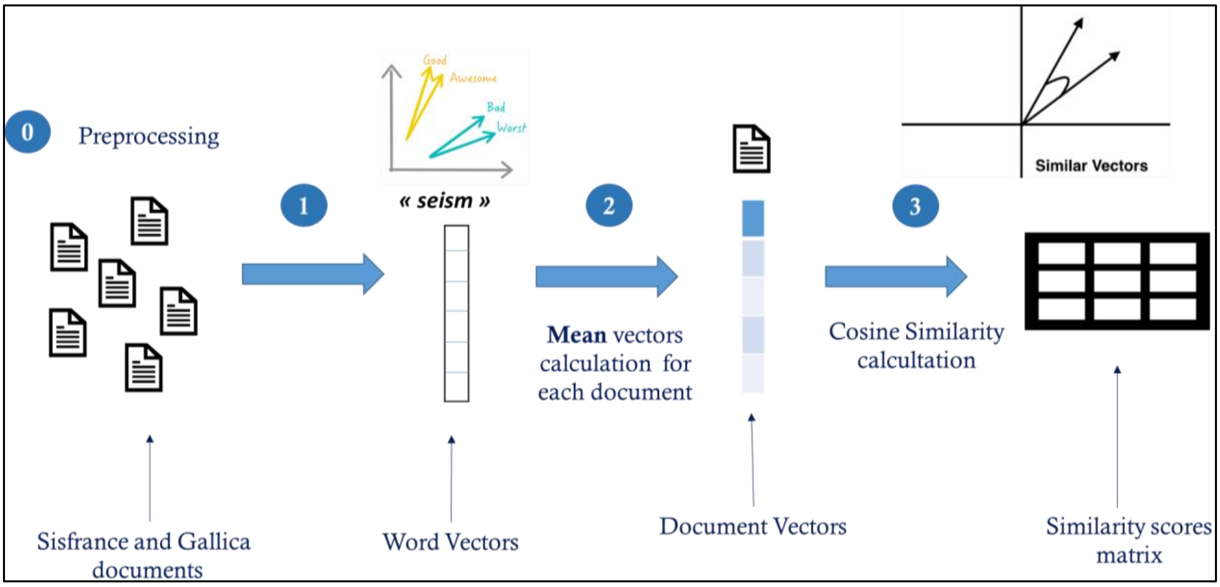


Figure 17 - Similarity process chain

Figure 18 illustrates the cosine similarity (θ) approach. The cosine looks at the angle between vectors; in this example SisFrance Document 2 is very similar to Gallica Document 1.

Based on the term weighting scheme, each document is represented by a vector of term weights

$$w(d) = w(d, w_1), w(d, w_2), \cdots, w(d, w_v)$$

The similarity between two documents d1 and d2 can be computed. One of the most widely similarity measures is cosine similarity computed as follow:

$$S(d_1, d_2) = \cos\theta = \frac{d_1 \cdot d_2}{\sqrt{\sum_{i=1}^{v} w_{1i}^2} \cdot \sqrt{\sum_{i=1}^{v} w_{2i}^2}}$$

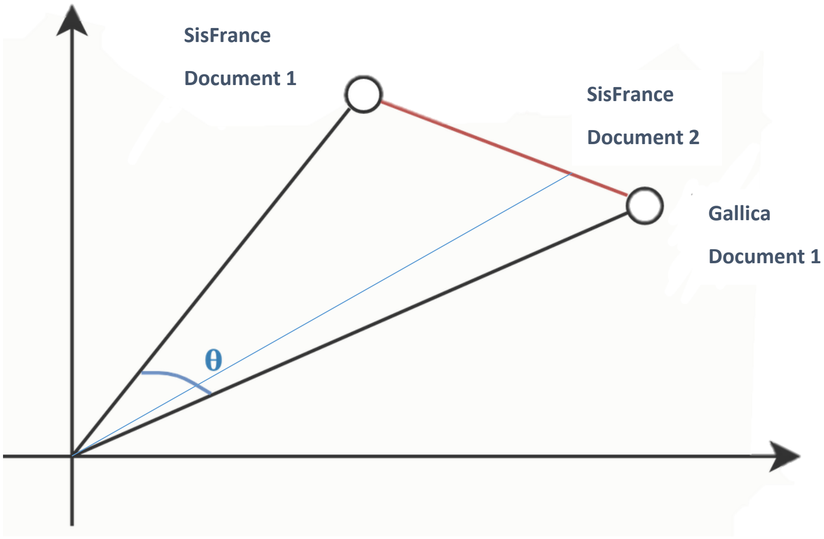
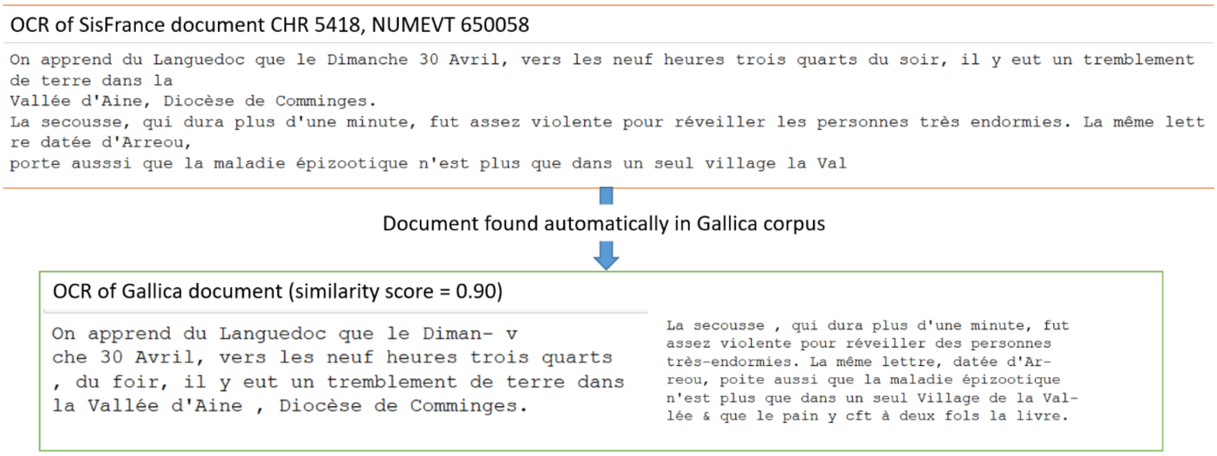


Figure 18 - Illustration of cosine similarity approach

Just below, an example of the similarity score obtained between a SisFrance document (CHR 5418, NUMEVT 650058) and an extract of a Gallica document.



OCR of SisFrance document CHR 5418, NUMEVT 650058

```
On apprend du Languedoc que le Dimanche 30 Avril, vers les neuf heures trois quarts du soir, il y eut un tremblement
de terre dans la
Vallée d'Aine, Diocèse de Comminges.
La secousse, qui dura plus d'une minute, fut assez violente pour réveiller les personnes très endormies. La même lett
re datée d'Arreou,
porte ausssi que la maladie épizootique n'est plus que dans un seul village la Val
```

Document found automatically in Gallica corpus

OCR of Gallica document (similarity score = 0.90)

```
On apprend du Languedoc que le Diman- v
che 30 Avril, vers les neuf heures trois quarts
, du foir, il y eut un tremblement de terre dans
la Vallée d'Aine , Diocèse de Comminges.
```

```
La secousse , qui dura plus d'une minute, fut
assez violente pour réveiller des personnes
très-endormies. La même lettre, datée d'Ar-
reou, poite aussi que la maladie épizootique
n'est plus que dans un seul Village de la Val-
lée & que le pain y cft à deux fols la livre.
```

Figure 19 - Similarity results, example

This methodology proves its efficiency in two ways:

Exclusion of known documents, already available in the SisFrance database, when the similarity score is very close to 1. In 90% of the cases, documents mentioning seisms felt in mainland France correspond either to a recording of a documents already available in the SisFrance database or do not include any new details. These documents can be removed automatically.

Find new documents dealing with earthquakes when the similarity score is greater than 0.65. The thematic is preserved.

The results of the similarity process will be discussed in the **section Results.**

5.4. Tools & IHM

Figure 20 sums up two available tools provided by QWAM: a dashboard and an Electronic Document Management System (EMDS) in order to respectively create requests and check original documents and display different graphics, and to ii) qualify documents: tag documents including information about seisms and documents which don't, let's call them noise.

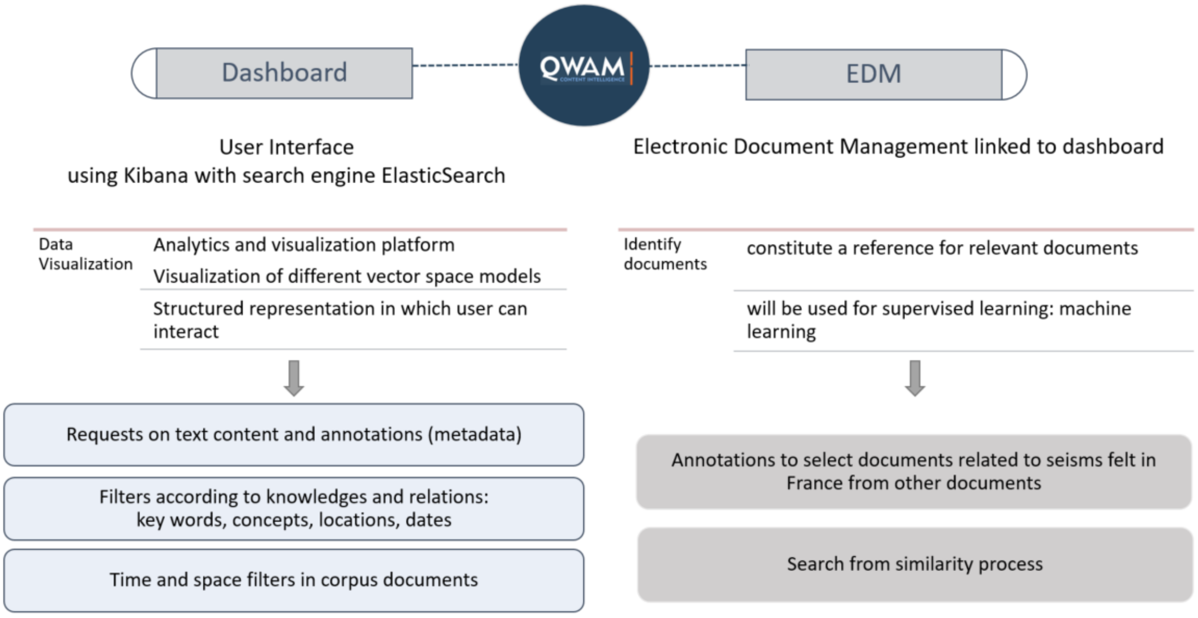

Dashboard
QWAM
EDM
User Interface
using Kibana with search engine ElasticSearch
Electronic Document Management linked to dashboard
Data Visualization
Analytics and visualization platform
Visualization of different vector space models
Structured representation in which user can interact
Identify documents
constitute a reference for relevant documents
will be used for supervised learning: machine learning
Requests on text content and annotations (metadata)
Filters according to knowledges and relations: key words, concepts, locations, dates
Time and space filters in corpus documents
Annotations to select documents related to seisms felt in France from other documents
Search from similarity process

Figure 20 - Materialization of data mining and document qualification

5.4.1. Dashboard

In order to realize search operations on a corpus of documents, QWAM provides a user interface (Kibana console, one of the main components of the Elastic Stack [24]), to ease document visualization and to display all available filtering applications.

This interface (**Figure 21**) can be considered as a portal between user and documents, allowing to pair documents from the database with user requests. This is possible thanks to the metadata enrichment and its indexation and annotation from previous steps:

- ✓ Seismological ontology from the SisFrance analysis,
- ✓ Contextual terms (locations, dates, events, organization),
- ✓ Metadata from the Gallica documents themselves,
- ✓ Identified relations between concepts.

All these metadata can thus be search criteria and are materialized as filters which can be overlaid, allowing for a multidimensional view of the requests.

This interface uses a dashboard to shape the data into interactive views.

- ✓ Requests on text content and annotations (metadata),
- ✓ Filters by knowledge and relations: key words, concepts, locations, dates,
- ✓ Time filtering with time line tool,
- ✓ Spatial filtering with interactive map.

Time and spatial filters will be useful to focus the search on a target past earthquake.

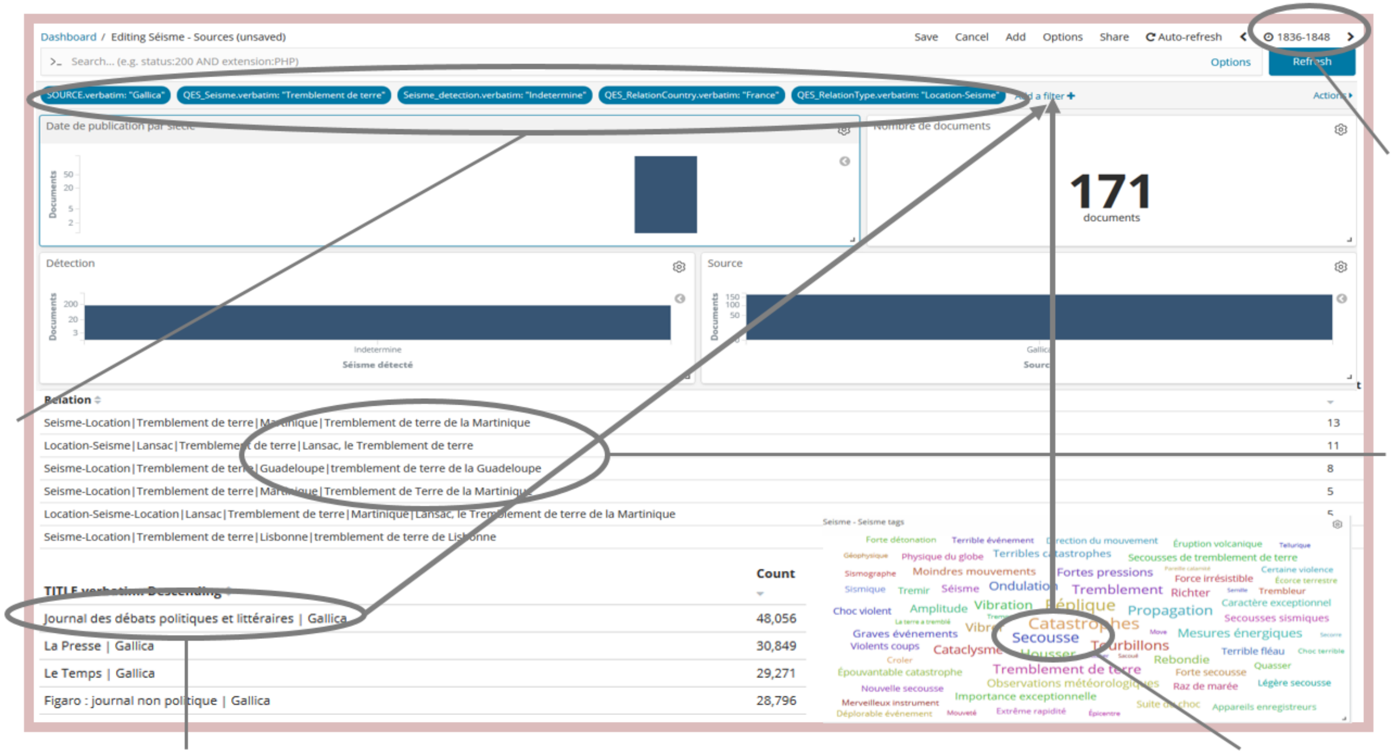
Filters applied

QES_Seisme exist

Relation type is **Seisme-Dégâts-Bati**

Relation country is **France**

Detection.verbatim is **Indetermined**



Dashboard / Editing Séisme - Sources (unsaved)

Search... (e.g. status:200 AND extension:PHP)

SOURCE.verbatim: "Gallica" | QES_Seisme.verbatim: "Tremblement de terre" | Seisme_detection.verbatim: "Indetermine" | QES_RelationCountry.verbatim: "France" | QES_RelationType.verbatim: "Location-Seisme"

Nombre de documents: 171 documents

Relation	
Seisme-Location|Tremblement de terre|Martinique|Tremblement de terre de la Martinique	13
Location-Seisme|Lansac|Tremblement de terre|Lansac, le Tremblement de terre	11
Seisme-Location|Tremblement de terre|Guadeloupe|tremblement de terre de la Guadeloupe	8
Seisme-Location|Tremblement de terre|Martinique|Tremblement de Terre de la Martinique	5
Location-Seisme-Location|Lansac|Tremblement de terre|Martinique|Lansac, le Tremblement de terre de la Martinique	5
Seisme-Location|Tremblement de terre|Lisbonne|tremblement de terre de Lisbonne	

TITLE.verbatim: Descending	Count
Journal des débats politiques et littéraires | Gallica	48,056
La Presse | Gallica	30,849
Le Temps | Gallica	29,271
Figaro : journal non politique | Gallica	28,796

Selection of historical period

Vocabulary is adapted according to chosen period

Verification of filter relevancy

Source document filter

Improving request by adding new keywords

Figure 21 - Dashboard (Kibana), search materialization

5.4.2. EMDS

Given the massive collection of harvested documents from the Gallica website, it is crucial to easily classify documents and this for two main reasons:

- ✓ Discriminate relevant documents from the other ones. What are relevant documents for this dedicated study? We want to retrieve documents relative to records on past earthquakes felt in Metropolitan France.
- ✓ Distinguish already processed documents from the other ones.

This classification is performed thanks to an Electronic Document Management System (EDMS), see **Figure 23**. It is a software system to organize and store different kinds of documents. This type of system is a very particular kind of document management system compared to a more general type of storage system that helps users to organize and store paper or digital documents. In order to provide good classification for digital documents, many electronic document management systems rely on a detailed process for document storage, including metadata.

Beyond practical aspect, this classification helps to set up supervised learning methods such as machine learning techniques (Artificial Intelligence, AI) pertaining to infer a function or to learn to classify from the training data in order to perform predictions on unseen data. This perspective won't be discussed in this work, as the implementation of such a method and its performance need a very large dataset of classified documents. This perspective could be considered in a second time.

Documents are thus divided into four categories:

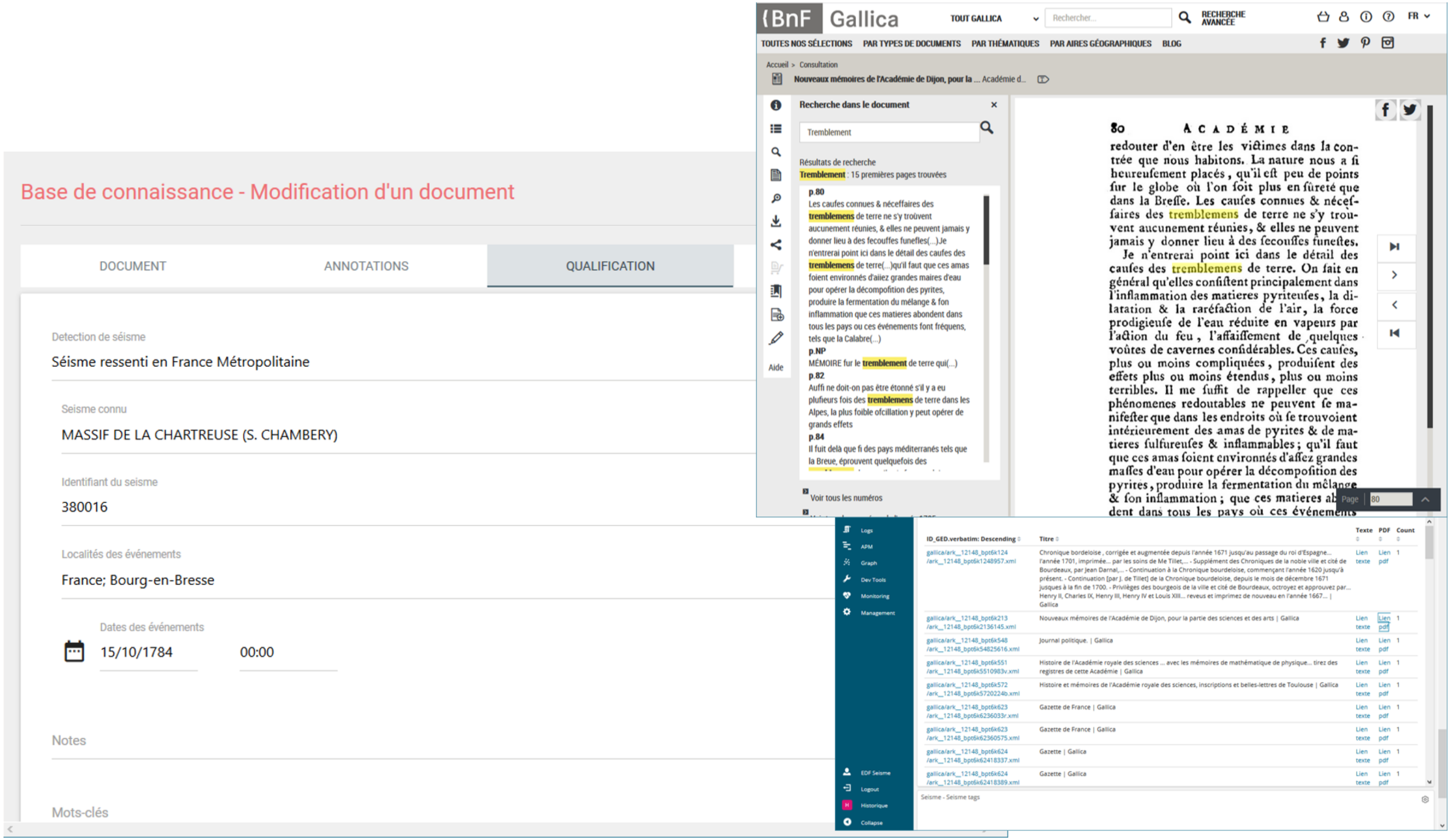
Indetermined	Document is not reviewed by user
Non Seism	Document does not include any information on past earthquakes
Seism not felt in Metropolitan France	Document includes information on past earthquakes but there is no information on effect felt in Metropolitan France
Seism felt in Metropolitan France	Document includes information on past felt in Metropolitan France

Figure 22 - Document classification un EMD System

For documents tagged as "seism felt in Metropolitan France", EDMS is configured to allow for adding fields to better characterize seisms and to link them (if it is possible) to earthquakes listed in the SisFrance database. Two fields are added:

- ✓ NUMEVT field (past earthquake ID from SisFrance database) to know if the earthquake mentioned in the document is already listed in the SisFrance database,
- ✓ CHRONO field (document ID from SisFrance database) to know if this document is already referenced in the SisFrance database.

The qualification of documents tagged as “non seism” or “seism not felt in Metropolitan France” will be used to raise frequent ambiguities and to teach the Artificial Intelligence to recognize false positives or ambiguities.



Base de connaissance - Modification d'un document

DOCUMENT | ANNOTATIONS | QUALIFICATION

Detection de séisme: Séisme ressenti en France Métropolitaine

Seisme connu: MASSIF DE LA CHARTREUSE (S. CHAMBERY)

Identifiant du seisme: 380016

Localités des événements: France; Bourg-en-Bresse

Dates des événements: 15/10/1784 00:00

Notes

Mots-clés

Nouveaux mémoires de l'Académie de Dijon, pour la ... Académie d...

Recherche dans le document

Tremblement

Résultats de recherche

Tremblement : 15 premières pages trouvées

p.80
Les caufes connues & néceffaires des tremblemens de terre ne s'y troùvent aucunement réunies, & elles ne peuvent jamais y donner lieu à des fecouffes funefles(...)Je n'entrerai point ici dans le détail des caufes des tremblemens de terre(...)qu'il faut que ces amas foient environnés d'aiiez grandes maires d'eau pour opérer la décompofition des pyrites, produire la fermentation du mélange & fon inflammation que ces matieres abondent dans tous les pays ou ces événements font fréquens, tels que la Calabre(...)

p.NP
MÉMOIRE fur le tremblement de terre qui(...)

p.82
Auffi ne doit-on pas être étonné s'il y a eu plufieurs fois des tremblemens de terre dans les Alpes, la plus foible ofcillation y peut opérer de grands effets

p.84
Il fuit delà que fi des pays méditerranés tels que la Breue, éprouvent quelquefois des

Voir tous les numéros

80 ACADÉMIE

redouter d'en être les victimes dans la contrée que nous habitons. La nature nous a ſi heureuſement placés, qu'il eſt peu de points ſur le globe où l'on ſoit plus en ſûreté que dans la Breſſe. Les cauſes connues & néceſſaires des tremblemens de terre ne s'y trouvent aucunement réunies, & elles ne peuvent jamais y donner lieu à des ſecouſſes funeſtes.

Je n'entrerai point ici dans le détail des cauſes des tremblemens de terre. On ſait en général qu'elles conſiſtent principalement dans l'inflammation des matieres pyriteuſes, la dilatation & la raréfaction de l'air, la force prodigieuſe de l'eau réduite en vapeurs par l'action du feu, l'affaiſſement de quelques voûtes de cavernes conſidérables. Ces cauſes, plus ou moins compliquées, produiſent des effets plus ou moins étendus, plus ou moins terribles. Il me ſuffit de rappeller que ces phénomenes redoutables ne peuvent ſe manifeſter que dans les endroits où ſe trouvoient intérieurement des amas de pyrites & de matieres ſulfureuſes & inflammables; qu'il faut que ces amas ſoient environnés d'aſſez grandes maſſes d'eau pour opérer la décompoſition des pyrites, produire la fermentation du mêlange & ſon inflammation; que ces matieres ab dent dans tous les pays où ces événements

Page 80

ID_GED.verbatim: Descending	Titre	Texte	PDF	Count
gallica/ark__12148_bpt6k124 /ark__12148_bpt6k1248957.xml	Chronique bordeloise , corrigée et augmentée depuis l'année 1671 jusqu'au passage du roi d'Espagne... l'année 1701, imprimée... par les soins de Me Tillet,... - Supplément des Chroniques de la noble ville et cité de Bourdeaux, par Jean Darnal,... - Continuation à la Chronique bourdeloise, commençant l'année 1620 jusqu'à présent. - Continuation [par J. de Tillet] de la Chronique bourdeloise, depuis le mois de décembre 1671 jusques à la fin de 1700. - Privilèges des bourgeois de la ville et cité de Bourdeaux, octroyez et approuvez par... Henry II, Charles IX, Henry III, Henry IV et Louis XIII... reveus et imprimez de nouveau en l'année 1667... | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k213 /ark__12148_bpt6k2136145.xml	Nouveaux mémoires de l'Académie de Dijon, pour la partie des sciences et des arts | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k548 /ark__12148_bpt6k54825616.xml	Journal politique. | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k551 /ark__12148_bpt6k5510983v.xml	Histoire de l'Académie royale des sciences ... avec les mémoires de mathématique de physique... tirez des registres de cette Académie | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k572 /ark__12148_bpt6k5720224b.xml	Histoire et mémoires de l'Académie royale des sciences, inscriptions et belles-lettres de Toulouse | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k623 /ark__12148_bpt6k6236033r.xml	Gazette de France | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k623 /ark__12148_bpt6k62360575.xml	Gazette de France | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k624 /ark__12148_bpt6k62418337.xml	Gazette | Gallica	Lien texte	Lien pdf	1
gallica/ark__12148_bpt6k624 /ark__12148_bpt6k62418389.xml	Gazette | Gallica	Lien texte	Lien pdf	1

Seisme - Seisme tags

Figure 23 – EMDS, qualification materialization

5.4.3. Back-Up

Gallica documents qualified through the EMD System are backed-up on server. All documents are thus available to be examined carefully by historians. These new records could populate in fine the SisFrance database and could help to revise IDP values or the epicentral intensity assigned to an earthquake.

6. Results

To validate the research system set up and described in the previous sections, a series of documentary reviews (called "qualification campaigns") were carried out. The first step of each qualification campaign is to manually classify documents according to the origin of the earthquake itself mentioned in the document (see 5.4.2 and Figure *22*). Each qualification campaign led to new relevant results for this project: unknown documents dealing with testimonies on past earthquakes occurred in mainland France. These documents were examined carefully, and all these results are presented in the next section.

Several campaigns were performed. Each campaign was subject to bring modification to the system: refine the precision strategy**. Figure 25** & **Figure 26** show all the campaigns carried out and parameter adjustments realized after each one.

6.1. Quantitative results

6550 documents are manually reviewed (qualified) from the Gallica collection. It represents 0.15% of the total amount of the harvested corpus. These documents are classified, and results are shown in **Figure** *24***.** 44% of these documents are dealing with earthquakes, most of them occurred in mainland France.

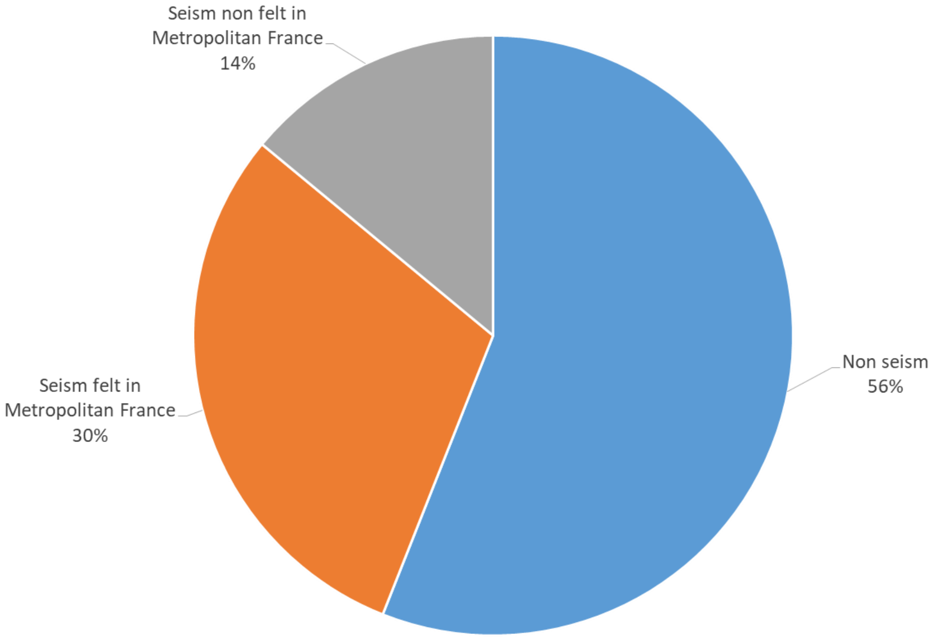
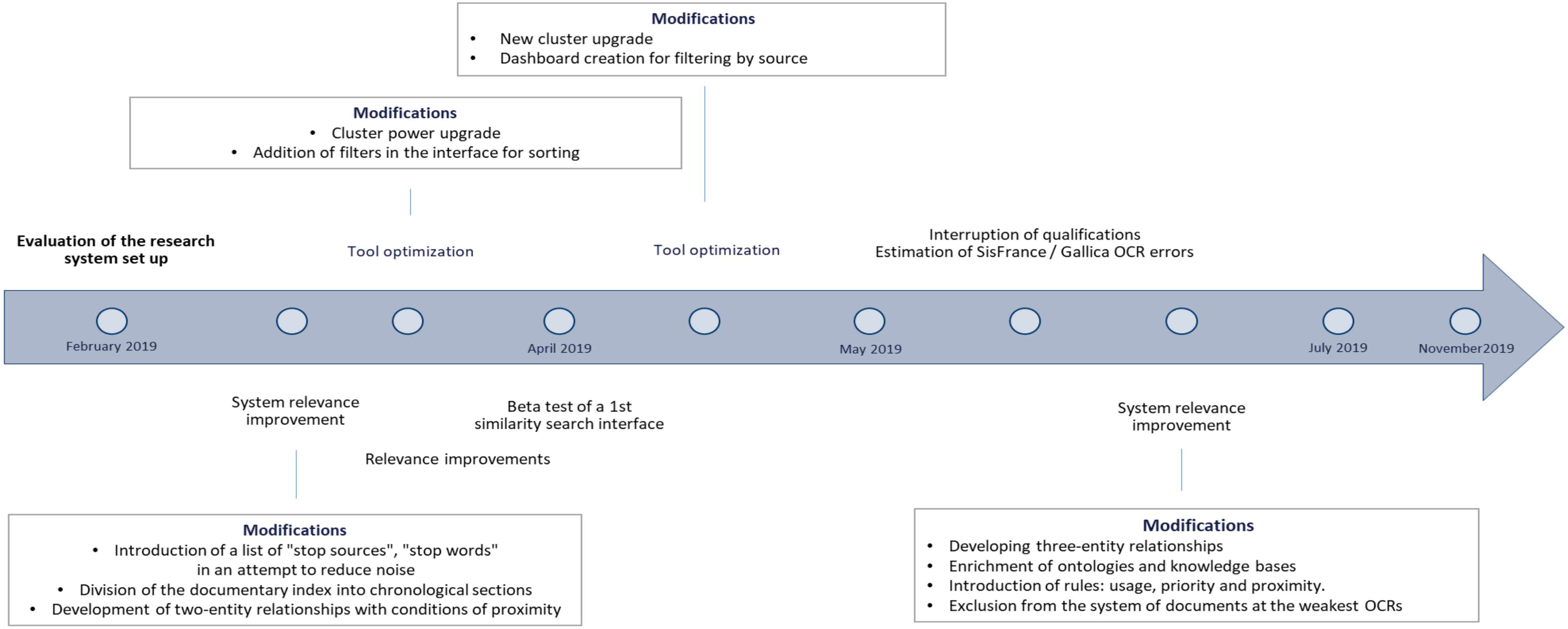


Figure 24 - Classification of 6550 documents manually screened



Modifications
- New cluster upgrade
- Dashboard creation for filtering by source

Modifications
- Cluster power upgrade
- Addition of filters in the interface for sorting

Evaluation of the research system set up

Tool optimization

Tool optimization

Interruption of qualifications
Estimation of SisFrance / Gallica OCR errors

System relevance improvement

Beta test of a 1st similarity search interface

System relevance improvement

Relevance improvements

Modifications
- Introduction of a list of "stop sources", "stop words" in an attempt to reduce noise
- Division of the documentary index into chronological sections
- Development of two-entity relationships with conditions of proximity

Modifications
- Developing three-entity relationships
- Enrichment of ontologies and knowledge bases
- Introduction of rules: usage, priority and proximity.
- Exclusion from the system of documents at the weakest OCRs

Figure 25 – 2019 Timeline - qualification campaigns

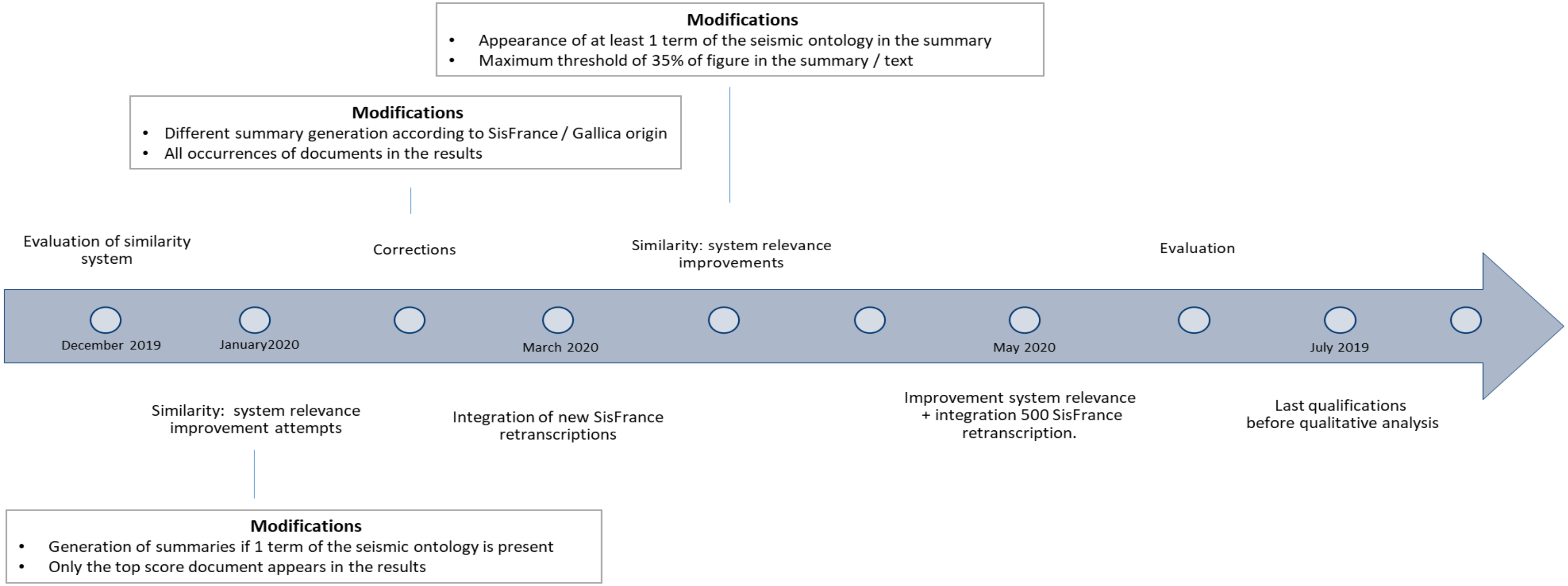

Modifications
• Appearance of at least 1 term of the seismic ontology in the summary
• Maximum threshold of 35% of figure in the summary / text
Modifications
• Different summary generation according to SisFrance / Gallica origin
• All occurrences of documents in the results
Evaluation of similarity system
Corrections
Similarity: system relevance improvements
Evaluation
December 2019
January2020
March 2020
May 2020
July 2019
Similarity: system relevance improvement attempts
Integration of new SisFrance retranscriptions
Improvement system relevance + integration 500 SisFrance retranscription.
Last qualifications before qualitative analysis
Modifications
• Generation of summaries if 1 term of the seismic ontology is present
• Only the top score document appears in the results

Figure 26 - 2019-2020 Timeline - qualification campaigns

The following section focuses on retrieved documents (total number: 1995) dealing with earthquakes felt in mainland France. First of all, the distribution of these documents is compared to the distribution of available documents in SisFrance (see **Figure 27**). Proportions are preserved.

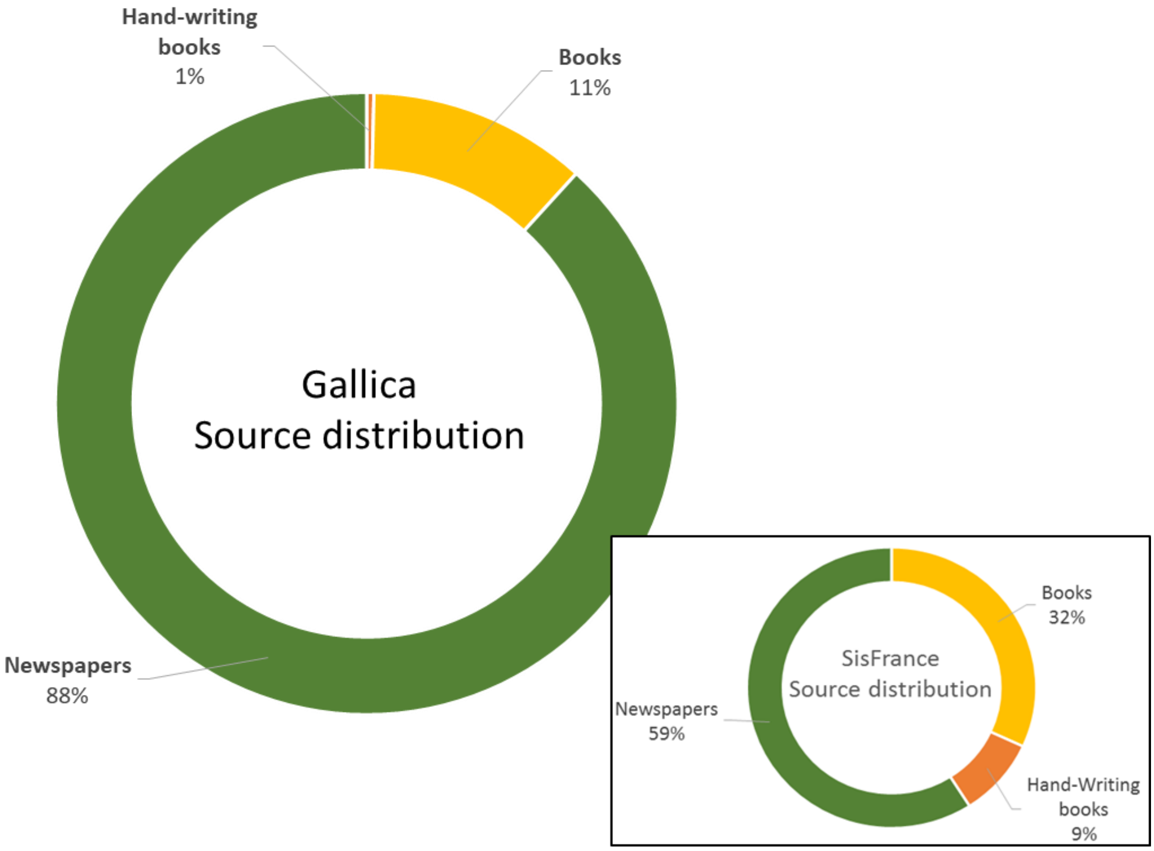


Figure 27 - Source distribution of the qualified documents “seism felt in Mainland France”

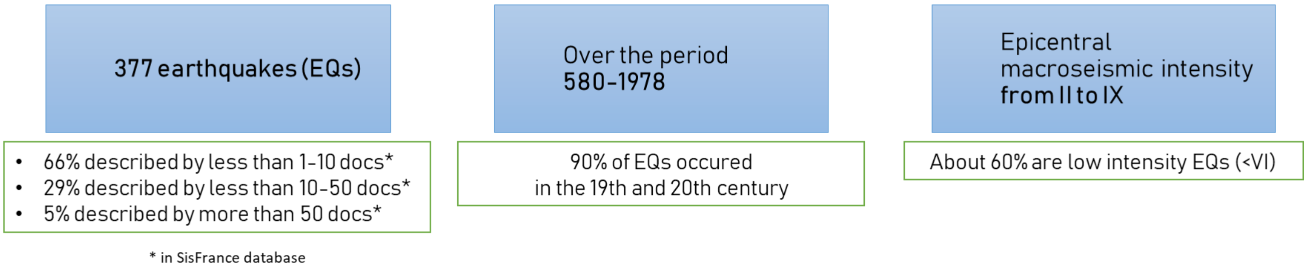
Analysis of ontologies for earthquakes

As shown previously in **section 4 Learning from existing database**, dedicated ontologies for earthquakes are determined. First of all, before analyzing the content of collected documents, a comparison is done between the occurrence of terms for each concept between the SisFrance documents and the Gallica documents dealing with earthquakes felt in mainland France. As expected, the frequency of appearance of words from the branch “seism” of seismological ontology is preserved. The coherency with other branches is less evident.

The frequency of appearance of words from SisFrance and Gallica documents is presented in **Appendix 2: Frequency of appearance of words from seismological ontology.** Results are presented for each branch of seismological ontology, using word clouds display.

387 documents referred to “unknown” earthquakes (EQs), in other words not referenced in the SisFrance 2017 database. Further analysis by historians will be necessary to determine if those are real earthquakes to be added to the SisFrance database or fake ones (explosion, fake story …). A list of unknown earthquakes and the links to documents are given in **Appendix 3: List of unknown earthquakes felt in mainland France.**

1608 documents could be linked to 377 EQs already known in the SisFrance 2017database. The following figure (**Figure 28**) summarizes all the details.



377 earthquakes (EQs)

- 66% described by less than 1-10 docs*
- 29% described by less than 10-50 docs*
- 5% described by more than 50 docs*

Over the period 580-1978

90% of EQs occured in the 19th and 20th century

Epicentral macroseismic intensity from II to IX

About 60% are low intensity EQs (<VI)

* in SisFrance database

Figure 28 – Characteristics of the qualified documents “seism felt in Metropolitan France”, referenced in SisFrance 2017 database

These 1608 documents are examined carefully in terms of content, by answering two questions:

(i) Does this document already exist in the SisFrance database (identical source: same title, date of publication and same content)?

(ii) If this document is not referenced in SisFrance (unknown document), does it provide new information and in fine, improve EQ knowledge?

The next figure (**Figure 29**) answers question (i). More than 60% of found documents are not referenced in the SisFrance 2017 database.

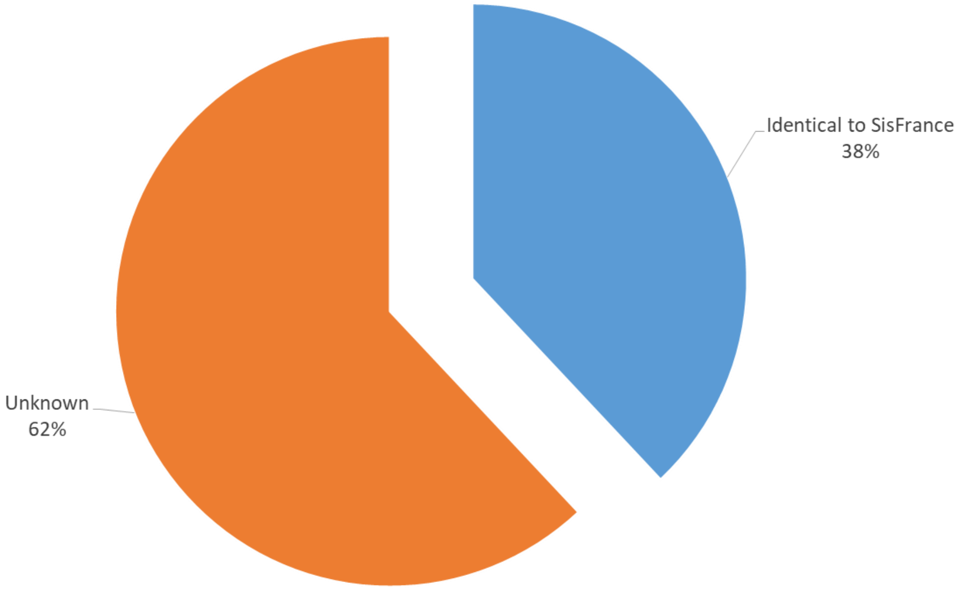


Figure 29 - Distribution of founded documents according to their content

To answer the second question, a qualitative analysis on the “unknown” documents is then performed to determine if these documents provide new information and improve knowledge of these earthquakes.

6.2. Qualitative results

To estimate the quality of these unknown documents from the SisFrance 2017 database, contents are examined very carefully. This work was done by a historian and a seismologist. This qualitative study focuses first on little known earthquakes referenced in the database, events with QPOS and QIE coefficient under or equal to C, but also on events described by less than 15 sources.

Finally 442 documents are first analyzed, dealing with 84 earthquakes poorly known in the SisFrance database. Among them, 39 documents are not referenced in the database (see **Figure 30**).

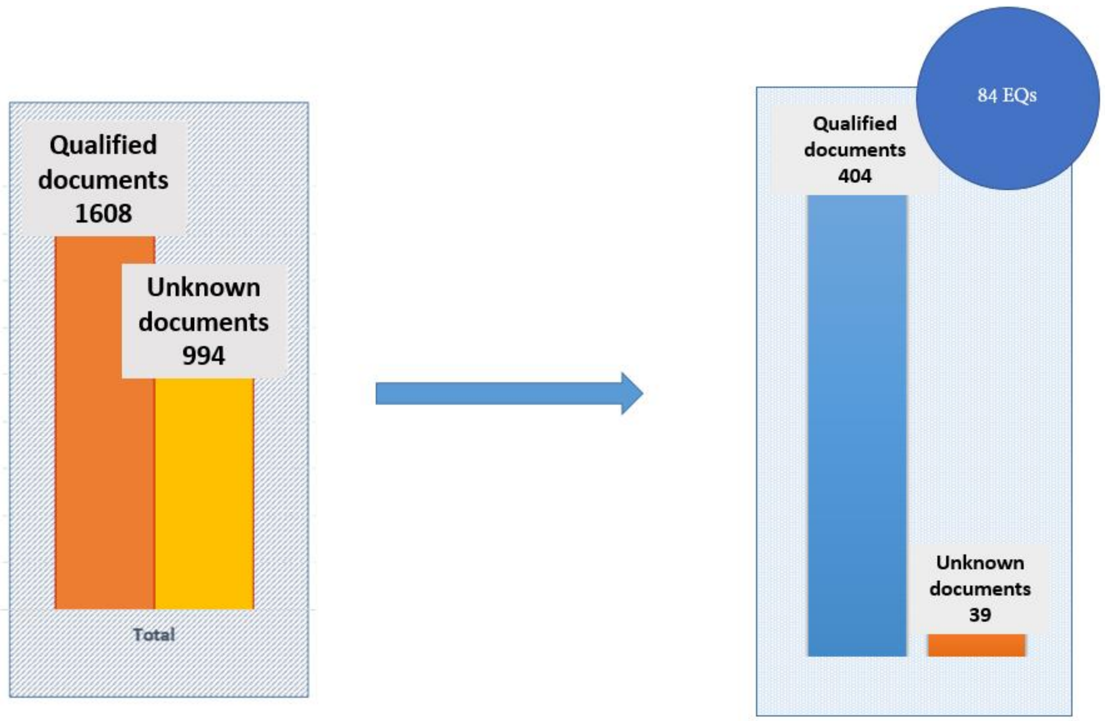


Figure 30 - Document selection for qualitative analysis after priority filtering

The followed qualitative analysis process is the same for each of documents (see **Figure 31**):

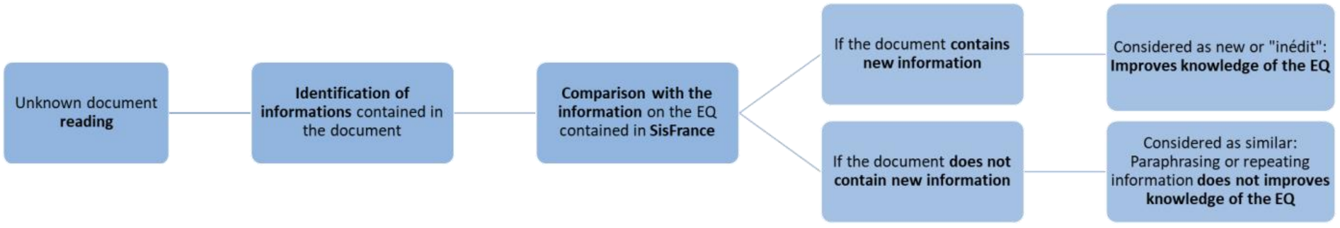

Unknown document reading
Identification of informations contained in the document
Comparison with the information on the EQ contained in SisFrance
If the document contains new information
If the document does not contain new information
Considered as new or "inédit": Improves knowledge of the EQ
Considered as similar: Paraphrasing or repeating information does not improves knowledge of the EQ

Figure 31 - Qualitative analysis process

This new information may concern:

- The location, duration and orientation of the tremors,
- The human (or animal) feeling of tremors,
- Human or structural damage.

The analysis led to new information, summarized in the table in **appendix 4**.

The found elements provided additional information by identifying new localities (IDP) with an intensity estimation assigned for each of them, or by revising the estimate in localities already listed in SisFrance. These new quotes will be sent to SisFrance in the form of a proposal associated with a reliability coefficient, the consortium must then approve them. All of these proposals are summarized in tableAppendix 5: Creation of observation points (IDPs) and proposed intensity value￼ andAppendix 6: Modification of observation points (IDPs) and proposed intensity value￼.

The next figure (**Figure 32**) sums up all results obtained for this first qualitative analysis. This new information will be submitted to the SisFrance consortium for approval.

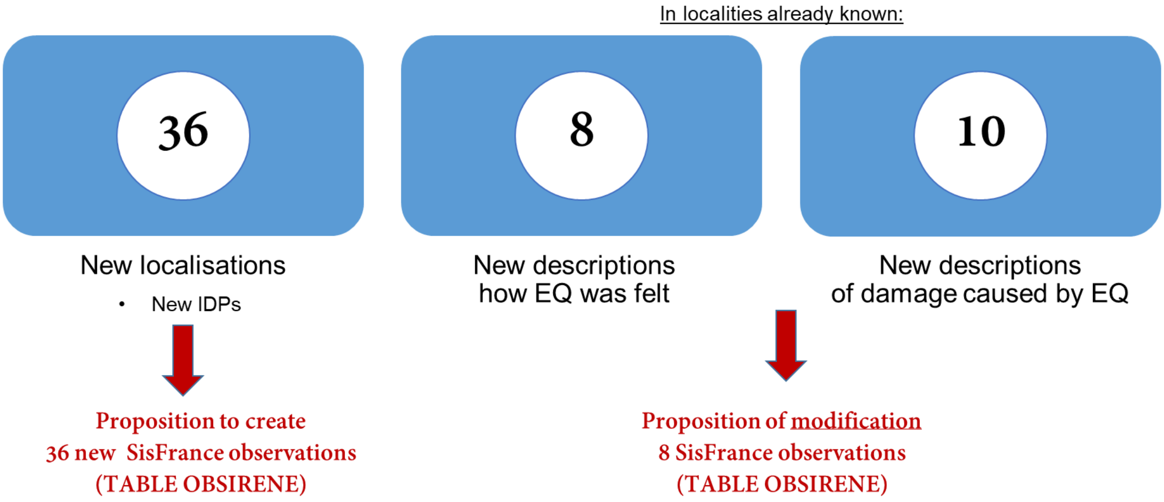

In localities already known:
36
8
10
New localisations
New IDPs
New descriptions how EQ was felt
New descriptions of damage caused by EQ
Proposition to create 36 new SisFrance observations (TABLE OBSIRENE)
Proposition of modification 8 SisFrance observations (TABLE OBSIRENE)

Figure 32 - Qualitative analysis, global results

6.3. Example of contextual analysis

In this section, some examples of the contextual analysis for new documents are given. To ease the reading, relevant passages have been highlighted according to a color code:

- Yellow: location,
- Green: sensations,
- Pink: duration of the tremor,
- Red: direction of the tremor,
- Blue: damages on people, buildings, furniture.

- **CEZALIER (BESLE) earthquake (SisFrance 2017 reference: NUMEVT = 630028)**

This EQ that occurred in October 1833 is described by 7 sources in the SisFrance database. Its epicentral intensity is fairly certain (QIE = B), whereas its epicentral location is very uncertain (QPOS =D), due to the few sources available:

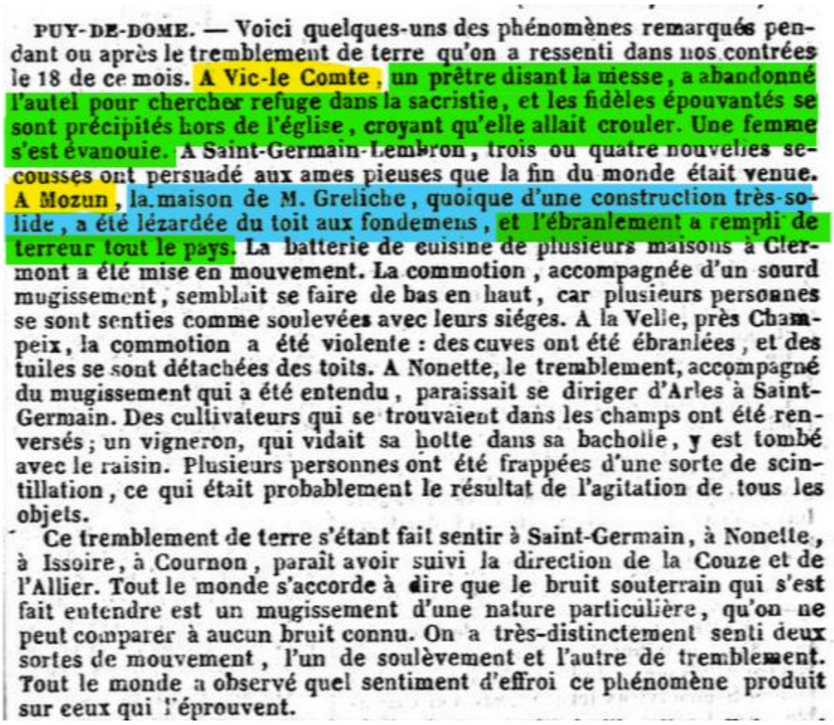
The data mining system allowed to retrieve a new document:

- *"Le constitutionnel", 26th October 1833 (Newspaper),* see **Figure 33**.

This document adds to our knowledge two new localities via reported testimonies of a priest in *Vic-le-Comte (Puy-de-Dôme)*, and of a resident in *Mauzun (Puy-de-Dôme)*.

The first gives details on the sensations of the shock felt by his parishioners. One of them faints.

The second relates that his house, even of solid construction, was cracked from the roof to the foundations.



PUY-DE-DOME. — Voici quelques-uns des phénomènes remarqués pendant ou après le tremblement de terre qu'on a ressenti dans nos contrées le 18 de ce mois. A Vic-le Comte, un prêtre disant la messe, a abandonné l'autel pour chercher refuge dans la sacristie, et les fidèles épouvantés se sont précipités hors de l'église, croyant qu'elle allait crouler. Une femme s'est évanouie. A Saint-Germain-Lembron, trois ou quatre nouvelles secousses ont persuadé aux ames pieuses que la fin du monde était venue. A Mozun, la maison de M. Greliche, quoique d'une construction très-solide, a été lézardée du toit aux fondemens, et l'ébranlement a rempli de terreur tout le pays. La batterie de cuisine de plusieurs maisons à Clermont a été mise en mouvement. La commotion, accompagnée d'un sourd mugissement, semblait se faire de bas en haut, car plusieurs personnes se sont senties comme soulevées avec leurs siéges. A la Velle, près Champeix, la commotion a été violente : des cuves ont été ébranlées, et des tuiles se sont détachées des toits. A Nonette, le tremblement, accompagné du mugissement qui a été entendu, paraissait se diriger d'Arles à Saint-Germain. Des cultivateurs qui se trouvaient dans les champs ont été renversés ; un vigneron, qui vidait sa hotte dans sa bacholle, y est tombé avec le raisin. Plusieurs personnes ont été frappées d'une sorte de scintillation, ce qui était probablement le résultat de l'agitation de tous les objets.

Ce tremblement de terre s'étant fait sentir à Saint-Germain, à Nonette, à Issoire, à Cournon, paraît avoir suivi la direction de la Couze et de l'Allier. Tout le monde s'accorde à dire que le bruit souterrain qui s'est fait entendre est un mugissement d'une nature particulière, qu'on ne peut comparer à aucun bruit connu. On a très-distinctement senti deux sortes de mouvement, l'un de soulèvement et l'autre de tremblement. Tout le monde a observé quel sentiment d'effroi ce phénomène produit sur ceux qui l'éprouvent.

Figure 33 - Extract of the article in "Le constitutionnel", published on October 1833

This allows us to add two new IDPs (to the 34 already known) as shown in the following **Figure 34**.

A quotation can be proposed to convert theses testimonies in a macroseismic intensity of, respectively, 5 (strong EQ) for IDP *Vic le Comte* and 7 (EQ with damage to buildings) for *Mauzun*.

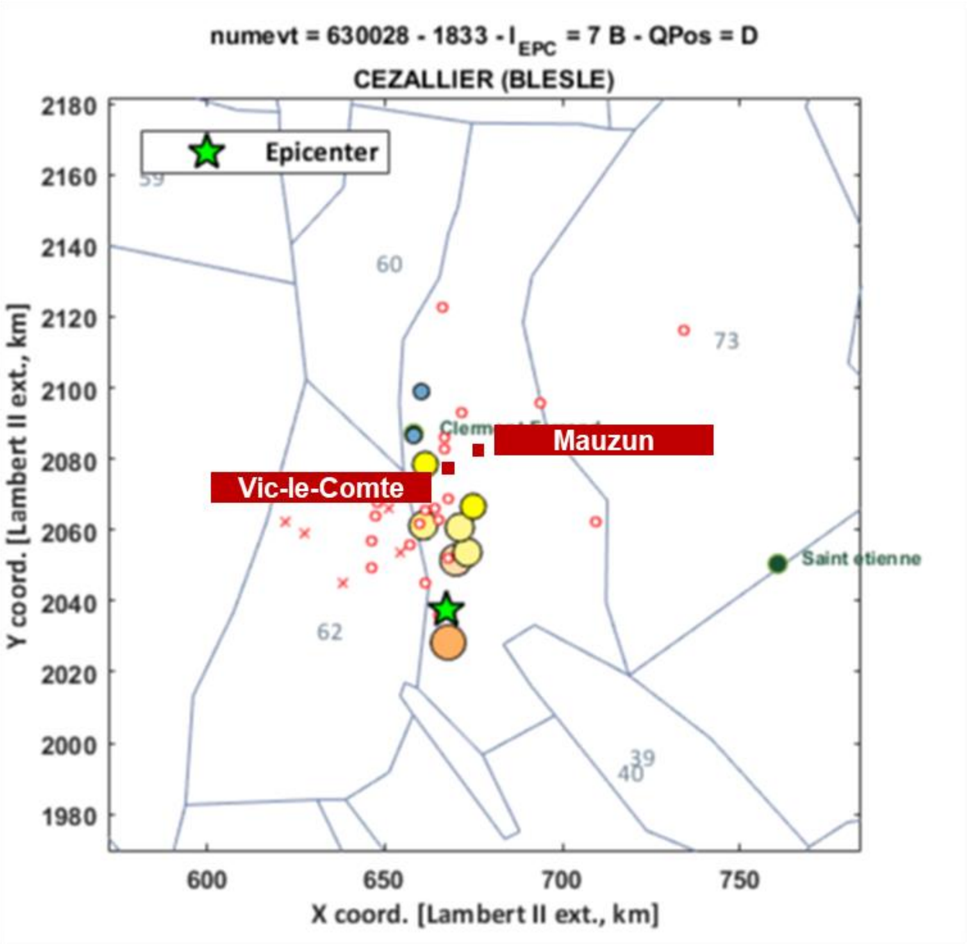


Figure 34 - Old and new IDPs on EQ 630028 Cezallier (Besle). For IDP scale, see Appendix 7: Intensity scale (IDP)

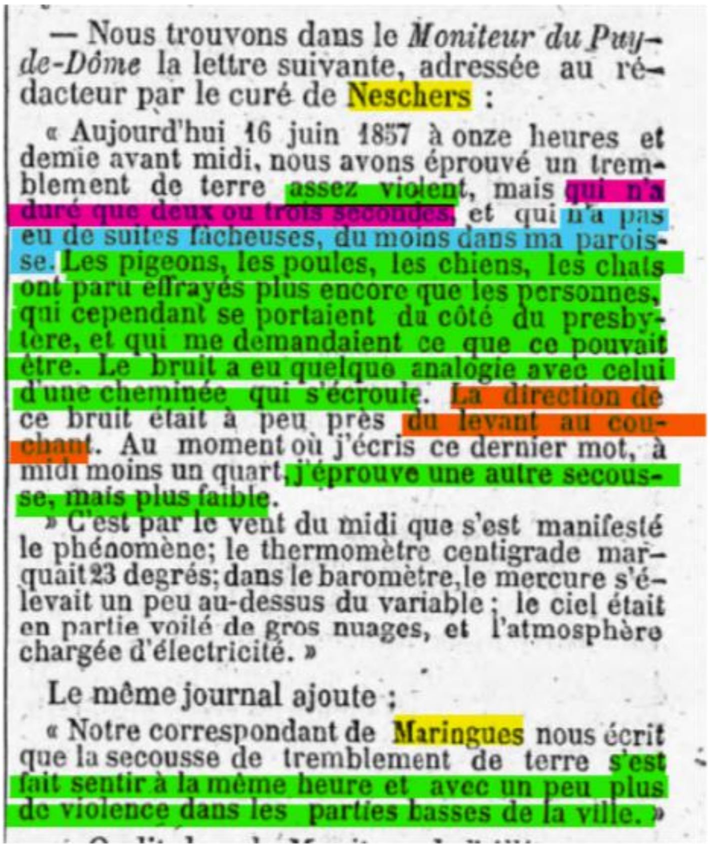
- **HAUTE-MARCHE (S. AUBUSSON?) earthquake (SisFrance 2017 reference: NUMEVT = 630042)**

This EQ that occurred in June 1857 is described by 20 sources in the SisFrance database. The reliability of its epicentral location and its epicentral intensity is very weak (QPOS, QIE = E).

New information can be found:

- *"La Presse", 20th June 1857 (Newspaper),* see **Figure 35**.

This document reports two new localities via a direct testimony of a priest in *Neschers (Loire)* and a reported testimony of a press correspondent in *Maringues* (*Puy-de-Dôme).* The priest described the sensations (humans and animals) and damages that the EQ caused in his town. The second part is about the feel of the tremor in the other town.



— Nous trouvons dans le *Moniteur du Puy-de-Dôme* la lettre suivante, adressée au rédacteur par le curé de Neschers :

« Aujourd'hui 16 juin 1857 à onze heures et demie avant midi, nous avons éprouvé un tremblement de terre assez violent, mais qui n'a duré que deux ou trois secondes, et qui n'a pas eu de suites fâcheuses, du moins dans ma paroisse. Les pigeons, les poules, les chiens, les chats ont paru effrayés plus encore que les personnes, qui cependant se portaient du côté du presbytère, et qui me demandaient ce que ce pouvait être. Le bruit a eu quelque analogie avec celui d'une cheminée qui s'écroule. La direction de ce bruit était à peu près du levant au couchant. Au moment où j'écris ce dernier mot, à midi moins un quart, j'éprouve une autre secousse, mais plus faible.

» C'est par le vent du midi que s'est manifesté le phénomène; le thermomètre centigrade marquait 23 degrés; dans le baromètre, le mercure s'élevait un peu au-dessus du variable; le ciel était en partie voilé de gros nuages, et l'atmosphère chargée d'électricité. »

Le même journal ajoute :

« Notre correspondant de Maringues nous écrit que la secousse de tremblement de terre s'est fait sentir à la même heure et avec un peu plus de violence dans les parties basses de la ville. »

Figure 35 - Extract of the article in ""Presse"", published in June 1857

This allows us to add two new IDPs in an area which was not covered by the 11 IDPs already known as shown in the following **Figure 36.**

A quotation can be proposed to convert theses testimonies in a macroseismic intensity of, respectively, 5 (strong EQ) for IDP *Neschers* and 3 (weakly felt EQ) for *Maringues*.

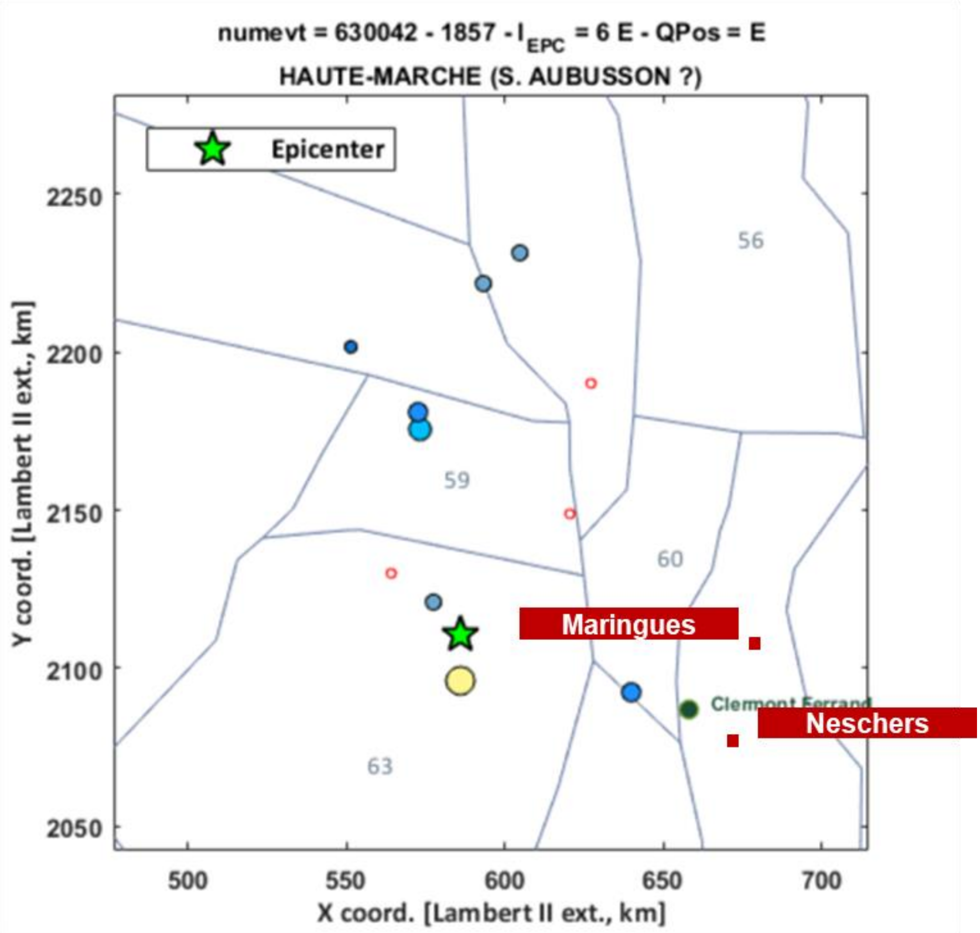

numevt = 630042 - 1857 - I$_{EPC}$ = 6 E - QPos = E
HAUTE-MARCHE (S. AUBUSSON ?)
Epicenter
56
59
60
63
Maringues
Neschers
Clermont-Ferrand
2250
2200
2150
2100
2050
Y coord. [Lambert II ext., km]
500
550
600
650
700
X coord. [Lambert II ext., km]

Figure 36 - Old and new IDPs on EQ 630042 Haute-Marche (S.Aubusson?). For IDP scale, see Appendix 7: Intensity scale (IDP)

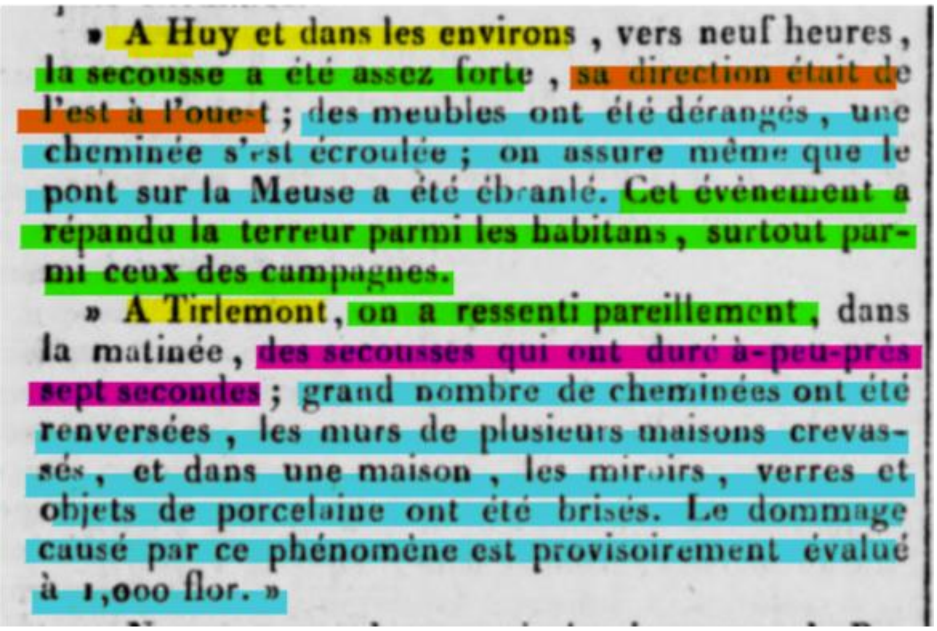
- **HAUTES-FAGNES earthquake (SisFrance 2017 reference: NUMEVT = 1100002)**

This EQ which occurred in December 1828 is described by 29 sources in the SisFrance database. According to the few sources available, and even if the QIE is fairly certain (B), the QPOS is weak: uncertain location (C).

New information can be found:

- *"Gazette Nationale ou Moniteur universel", 28th February 1857 (Newspaper).* See **Figure 37**.

This document reports two new localities via a reported testimony in *Huy* and *Tirlemont* (*Belgium).* The document described various information: direction and duration of the tremor, the sensations and damages that the EQ caused in these towns. Even a big bridge was shaken by the earthquake. We can converse theses testimonies in a macroseismic intensity of respectively 7 and 7.5 (EQ with damage to buildings).



» A Huy et dans les environs, vers neuf heures, la secousse a été assez forte, sa direction était de l'est à l'ouest ; des meubles ont été dérangés, une cheminée s'est écroulée ; on assure même que le pont sur la Meuse a été ébranlé. Cet événement a répandu la terreur parmi les habitans, surtout parmi ceux des campagnes.

» A Tirlemont, on a ressenti pareillement, dans la matinée, des secousses qui ont duré à-peu-près sept secondes ; grand nombre de cheminées ont été renversées, les murs de plusieurs maisons crevassés, et dans une maison, les miroirs, verres et objets de porcelaine ont été brisés. Le dommage causé par ce phénomène est provisoirement évalué à 1,000 flor. »

Figure 37 - Extract of the article in" Gazette Nationale ou Moniteur Universel", published in February 1857

This adds two new IDPs in an area which was not covered by the 36 IDPs already known as shown in the following **Figure 38**.

A quotation can be proposed to convert theses testimonies in a macroseismic intensity of, respectively, 7 (strong EQ) for IDP *Huy* and 7.5 (EQ with damage to buildings) for *Tirlemont*.

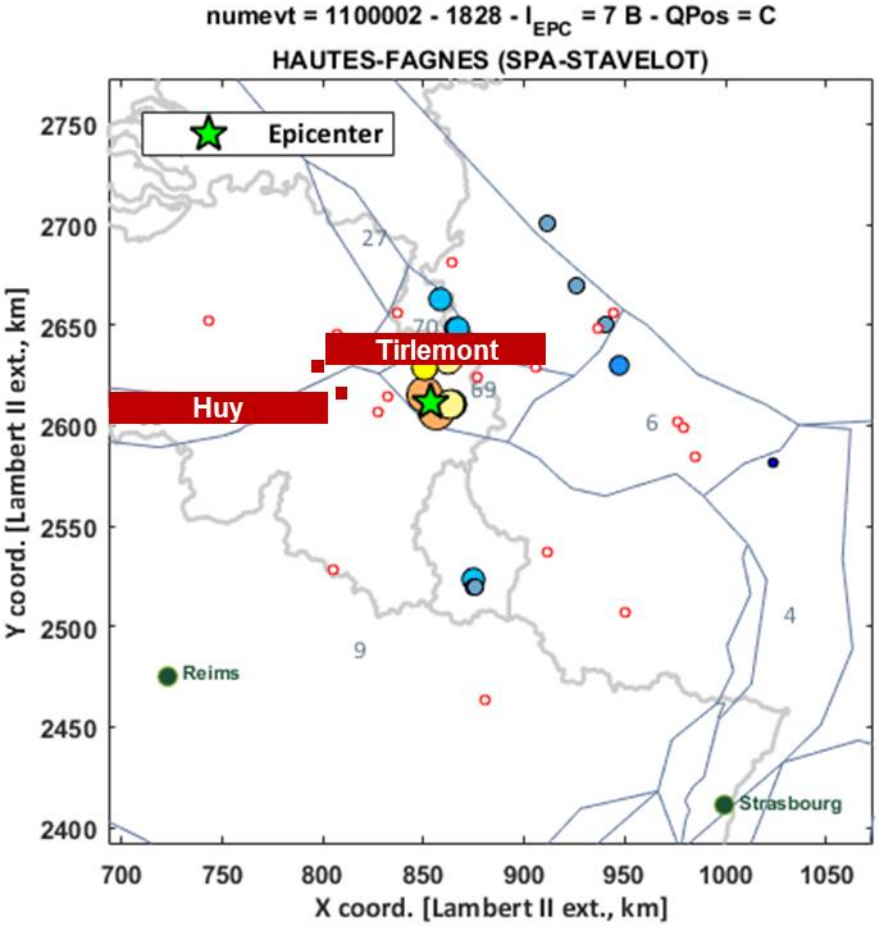

numevt = 1100002 - 1828 - I_EPC = 7 B - QPos = C
HAUTES-FAGNES (SPA-STAVELOT)
Epicenter
Tirlemont
Huy
Reims
Strasbourg
27
70
69
6
9
4
2750
2700
2650
2600
2550
2500
2450
2400
Y coord. [Lambert II ext., km]
700
750
800
850
900
950
1000
1050
X coord. [Lambert II ext., km]

Figure 38 - Old and new IDPs on EQ 1100002 Hautes-Fagnes (Spa-Stavelot). For IDP scale, see Appendix 7: Intensity scale (IDP)

As a conclusion, the datamining method research set up is efficient and allows to retrieve new testimonies to improve historical earthquakes knowledge.

First results are very promising, and a qualitative analysis will be done on the rest of retrieved documents (see **Figure 30**)

The next step is dedicated to increase the value of those new documents, by:

- Integrating them into the SisFrance database once they are approved by the SisFrance consortium,
- Contextualizing new records and submit them to historical and technical expertise, especially for the oldest ones.

7. Perspectives

Pursue research

- **Improve data mining techniques (BERT, CamemBERT and FlauBert)**

BERT (Bidirectional Encoder Representations from Transformers) is a recent paper [25] published by researchers at Google AI Language. BERT's key technical innovation is applying the bidirectional training of Transformer [26], a popular attention model, to language modelling. This contrasts with previous efforts which looked at a text sequence either from left to right or combined left-to-right and right-to-left training.

Unlike directional models, such as the Bag of Words model, which reads the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all its surroundings (left and right of the word). This is a very promising technique to find new records.

BERT contains multi-lingual models available but is not trained specifically for French-language corpora. New models such as FlauBERT [27] or CamemBERT [28] are now available which make use of large pre-trained models that capture specificities of the French language.

The chart below **Figure 39** is an illustration of the BERT process.

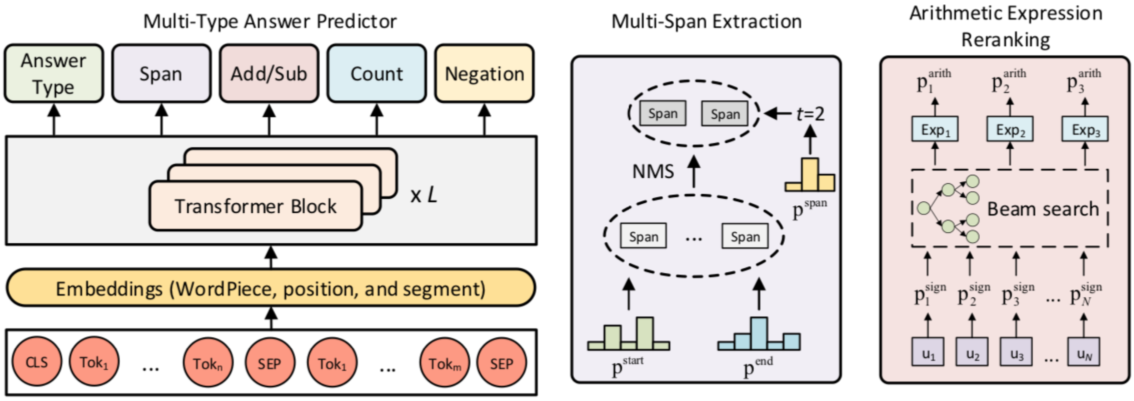
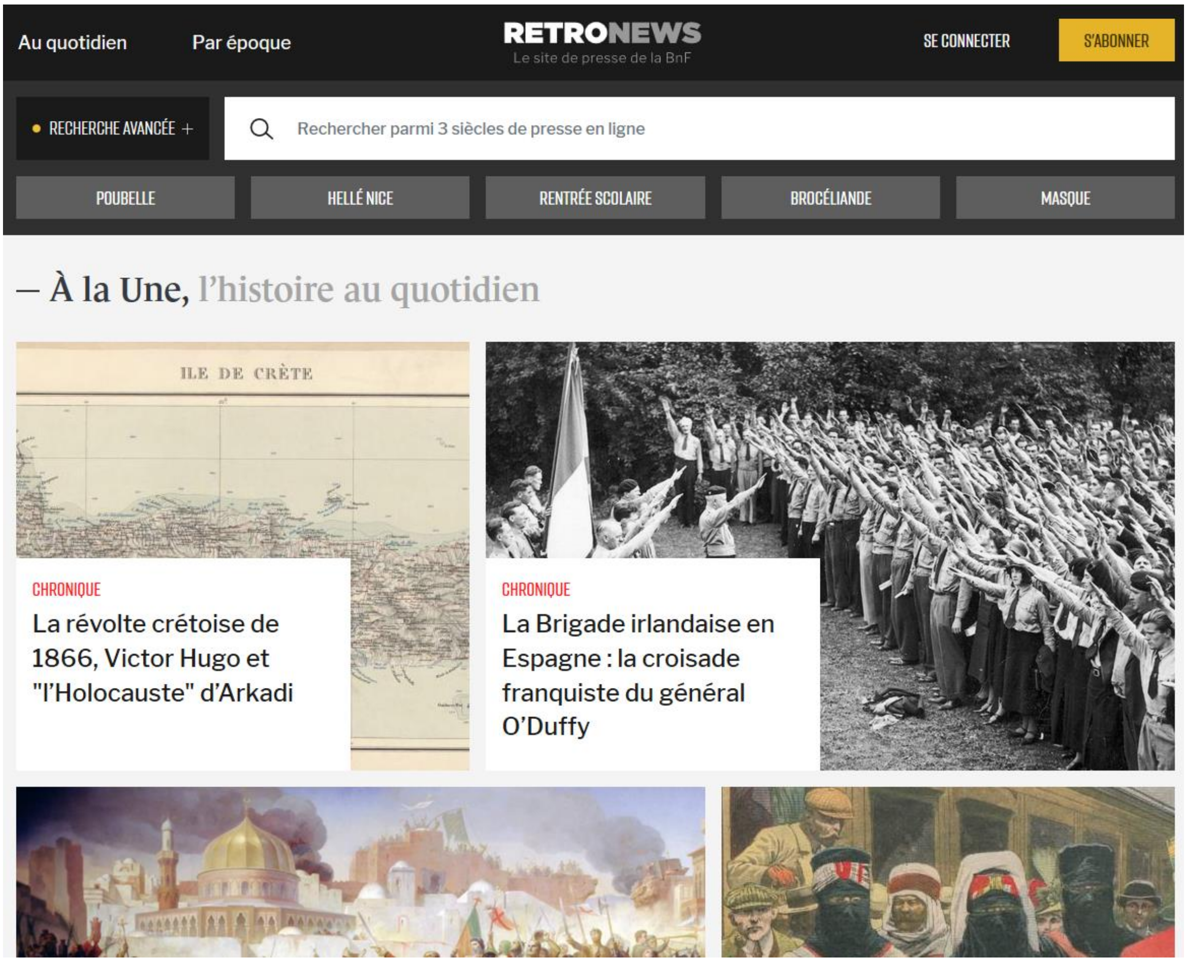


Figure 39 - BERT illustration (Full Credit: https://github.com/tomohideshibata/BERT-related-papers)

- **Look into other online databases: RetroNews**

RetroNews ([10], [11]) is a web platform created by the Bibliothèque nationale de France (BnF). It gives access to newspapers published in France between 1631 and 1966. More than 2 million documents are available. The same strategy that was applied to the Gallica collection will be retained to investigate this new one. Gallica and RetroNews websites are both from the BnF, the website structures are slightly identical, which facilitates the harvest of documents.



— À la Une, l'histoire au quotidien

CHRONIQUE

La révolte crétoise de 1866, Victor Hugo et "l'Holocauste" d'Arkadi

CHRONIQUE

La Brigade irlandaise en Espagne : la croisade franquiste du général O'Duffy

Figure 40 - RetroNews website ([10],[11])

Enhance methodology to other natural hazards

This built-in methodology proves its efficiency to find new records to improve past earthquake knowledge. As explained in this work, a system was built allowing for the exploitation of massive collection of documents. One of the key steps is the creation of seismological ontology used as dedicated dictionary to extract relevant information from the Gallica collection of documents.

This methodology can thus be applied to other natural hazards such as floods, windstorms, wave storms, heat waves, droughts, landslides), using dedicated ontology.

8. Conclusion

In this study, we implemented a method based on data mining techniques to improve historical seismicity knowledge by finding new records on literary heritage available on the web. This current work is focused on a specified available corpus of documents: Gallica, the digital library of the Bibliothèque nationale de France (BnF). Defining a precision strategy wouldn't be possible without contributions of several disciplines such as seismology, linguistic science, and computer science and historical.

The main part of this work focuses on designing methods and algorithms in order to effectively process more than 3.8 million documents harvested from Gallica, and to find relevant texts (records on past earthquakes felt in mainland France) through the background 'noise' (all other documents).

The success of this work is based on three key points:

- **Exploiting the SisFrance database** to define seismological ontology which is used as dedicated dictionary to extract relevant information from the Gallica collection of documents.
- **Information Extraction from Gallica corpus including** two fundamental tasks: entity recognition (dedicated ontology and classical named entity) and relation extraction.
- **Using advanced techniques of data mining:** especially the use of similarity process which dramatically helps to increase the number of records on past earthquakes felt in mainland France.

Up to now, more than 1600 documents dealing with earthquakes felt in mainland France have been found in the Gallica Corpus, and 62% of them are not listed in the SisFrance 2017 database. A qualitative analysis of documents dealing with little known earthquakes (QPOS < B or QIE < B) is performed. New IDP, new details on perception are discovered.

This new information will soon be given to the SisFrance consortium and will be compared to previously existing documents to reevaluate the characteristics of the events.

These first results are very promising as more than 60,000 pertinent documents from the Gallica collection are still unexplored. A new campaign will be performed using new advanced data mining techniques such as the BERT process to explore unseen documents.

The next objective would be the exploration of other databases available on other websites with this methodology. The RetroNews website seems to be the first valuable candidate as this platform is very similar to the one from Gallica (identical website architecture), and its collection of documents is very important: more than 2.5 million documents available.

This built-in methodology proves its efficiency to find new records to improve past earthquake knowledge. The next challenge will be to apply this methodology to other natural hazards such as floods, windstorms, wave storms, heat waves, droughts, landslides, using dedicated ontology.

List of figures

References

Seismological References

[1] Nocquet J-M, Calais E (2004) Geodetic measurements of crustal deformation in the Western Mediterranean and Europe. Pure Appl Geophys 161:661–681

[2] Walpersdorf A, Baize S, Calais E et al (2006) Deformation in the Jura Mountains (France): First results from semi-permanent GPS measurements. Earth Planet Sci Lett 245:365–372

[3] Manchuel, K., Traversa, P., Baumont, D., Cara, M., Nayman, E., & Durouchoux, C. (2018). The French seismic CATalogue (FCAT-17). Bulletin of Earthquake Engineering, 16, 2227-2251.

[4] Cara M, Cansi Y, Schlupp A et al (2015) Si-Hex: a new catalogue of instrumental seismicity for metropolitan France. Bull Soc Géol Fr 186:3–19. doi:10.2113/qssqfbull.186.1.3

[5] Medvedev SP, Sponheuer W, Karnik V (1967) Seismic intensity scale version 1964. Inst. Geody. Publ., Jena, p 48

[6] Lambert, J., Montfort-Climent, D., & Bouc, O. (2015). Catalogue of isoseismal areas for XXth century french historical earthquakes (Io > VI). Tech. rep., BRGM.

[7] SISFRANCE, https://sisfrance.irsn.fr/

Gallica & RetroNews

[8] Stanford Prize for Innovation in Research Libraries (SPIRL): Application from the Bibliothèque nationale de France (BnF) for Gallica (gallica.bnf.fr) and Data (data.bnf.fr), 2012. En ligne: https://library.stanford.edu/sites/default/files/Bibliotheque%20nationale%20de%20France.pdf

[9] Gallica Presse et revues : http://gallica.bnf.fr/html/und/presse-et-revues/presse-et-revues

[10] RetroNews: http://www.retronews.fr

[11] Blog RetroNews: http://blog.retronews.fr

Data Mining References

[12] [general] Baeza-Yates, R. & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley.

[13] [general] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases".http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdfRetrieved 2008-12-17.

[14] [general] Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery", International Journal of Information Technology and Decision Making, Volume 7, Issue 47: 639 –682. doi:10.1142/S0219622008003204

[15] [general] Solka J. L (2008) Text Data Mining: Theory and Methods, Statistic Surveys, Vol. 2 (2008) 94-112, doi: 10.214/07-SS016

[16] [crawl] C. Chen, Structuring and visualising the world-wide web with generalised similarity analysis. In: Proceedings of the 8th ACM Conference on Hypertext (Hypertext '97), Southampton, UK (April 1997). Available from: www. brunel.ac.uk/~cssrccc2/papers/ht97.pdf

[17] [crawl] Yong-Bin Yu, Shi-Lei Huang, Nyima Tashi, Huan Zhang, Fei Lei, Lin-Yang Wu. A Survey about Algorithms Utilized by Focused Web Crawler. Journal of Electronic Science and Technology, 2018, 16(2): 129-138

[18] [ocr] Sahu, Narendra & Sonkusare, Manoj. (2017). A Study on Optical Character Recognition Techniques. International Journal of Computational Science, Information Technology and Control Engineering. 4. 01-15. 10.5121/ijcsitce.2017.4101.

[19] [ocr] Rice Stephen V., Frank R. Jenkins, and Thomas A. Nartker The Fourth Annual Test of OCR Accuracy, expervision.com, retrieved 21 May 2013.

[20] [zipf's rule] Blanchard, A. Understanding and customizing stopword lists for enhanced patent mapping.World Patent Information, Elsevier, 2007, 29 (4), pp.308. 10.1016/j.wpi.2007.02.002. hal-01247971

[21] [word2vec] Mikolov, T. et al. (2013). "Efficient Estimation of Word Representations in Vector Space".

[22] [BoW] Zhao, Rui & Mao, Kezhi. (2017). Fuzzy Bag-of-Words Model for Document Representation. IEEE Transactions on Fuzzy Systems. PP. 1-1. 10.1109/TFUZZ.2017.2690222.

[23] [TF-IDF] Kim, S., Gil, J. Research paper classification systems based on TF-IDF and LDA schemes. Hum. Cent. Comput. Inf. Sci. 9, 30 (2019). https://doi.org/10.1186/s13673-019-0192-7

[24] [elasticsearch] https://www.elastic.co/

[25] [BERT] Devlin, J., Chang, M-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Biredirectional Transformers for Language Understanding, arXiv:1810.04805v2

[26] [TRANSFORMER] Vaswani, A., etal (2017). Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., arXiv:1706.03762v5

[27] [FlauBERT] Le, Hang & Vial, Loïc & Frej, Jibril & Segonne, Vincent & Coavoux, Maximin & Lecouteux, Benjamin & Allauzen, Alexandre & Crabbé, Benoît & Besacier, Laurent & Schwab, Didier. (2019). FlauBERT: Unsupervised Language Model Pre-training for French.

[28] [CamemBERT] Martin, Louis et al. “CamemBERT: a Tasty French Language Model.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020): n. pag. Crossref. Web.

Appendix 1: Seismological Ontology

ASSEMBLY CONCEPT

eglife, établissement, escalier, atelier, immeuble, lustre, sommier, banc, sacristie, salle, caserne, armoire, lampe, maison, château, presbytère, chambranle, donjon, tunnel, nef, poudrière, cabinet, refectoire, beffroi, cimetière, échafaudage, galerie, plancher, mobilier, demeure, perchoir, cave, chambre, trottoir, chaudron, temple, portail, bibliothèque, pavé, chapiteau, voute, abbaye, appartement, arcade, ardoise, balcon, basilique, bâtiment, bâtisse, béton, bloc de pierre, bloc de pierre détaché, brique, cabane, carreau, cathédrale, charpente, cheminée, clocher, clocheton, colonne, construction, corniche, démolir, écurie, édifice, église, étable, étage, extérieur, façade, fauteuil, fendu, fenêtre, fontaine, forêt, fossé, grange, grenier, habitation, hangar, les pots ont dansé dans les cuvettes, lit, logis, lustre, maison, moellon, monastère, monument, mortier, mortier tombant des cheminées et plafond, moulin, muraille, murs, paroisse, parquet, pavillon, pendule, pièce, plafond, plancher, plinthe, porte, prieuré, remuer les meubles dans les maisons, réparation, ruine, séparation visible des joints des charpentes et cloisons, siège, sous-sol, toit, toiture, tour, tuiles, villa, voûte.

DAMAGE CONCEPT

affaisser, balancement, balancer, blesser, blessure, bris, chute, chuter, commotion, craquement, craquer, décéder, décès, décombre, déplacer, désastre, destruction, dommage, éboulement, ébouler, éboulis, ébranlement, ébranler, endommager, enseveli, entraille, fente, fissure, fracasser, frémissement des vitres, gravité, heurter, lézarder, mort, oscillation, osciller, ravager, remuer, renverser, rupture, s'entrechoquer, tassement, tasser, tomber, vaciller, victime, affaissement, éclat, effondrement, amortissement, renversement, tintement, malheur, sinistre, dégâts, dégàts, écroulement, agitation, lézarde, désordre, écrouler, glissement, débordement, excavation, chanceler, frémir, respapé, survivant, sinistré.

SEISMIC CONCEPT

amplitude, cataclysme, catastrophe, convulsion terrestre, déchirement, déchirer, degré, écorce terrestre, épicentre, force, foyer, géophysique, intense, intensité, macroseismique, magnitude, mouvement, propagation, richter, ritcher, secousse, séisme, seismique, sismique, sismogène, sismographe, tellurique, tremblement, trembleterre, tremblotement, trémolo, vibration, vibrer, oscillation, agitation, ondulation, désastre, sinistre, cataclisme, crise, tourbillons, hypocentre, séismologie, violent, élevé, prononcé, modéré, épicentral, macrosismique, pléistoseiste, macroseis, , réplique, séismographe.

BEHAVIOR CONCEPT

abattu, acourir, affliction, affolement, affoler, agitation, agiter, alarme, bouleverser, bousculade, chahut, choc, craindre, crainte, déranger, désarroi, désolation, effrayer, effroi, émoi, enfuir, épouvanter, éprouver, étourdir, étourdissement, éveiller, frayeur, frissonnement, inquiéter, malaise, palpitation, réveiller, réveiller en sursaut, s'échapper sous leurs pieds, secours, s'émouvoir, sensible, soubresaut, stupéfaction, surprise, témoignage, trépidation, trésaillement.

NOISE CONCEPT

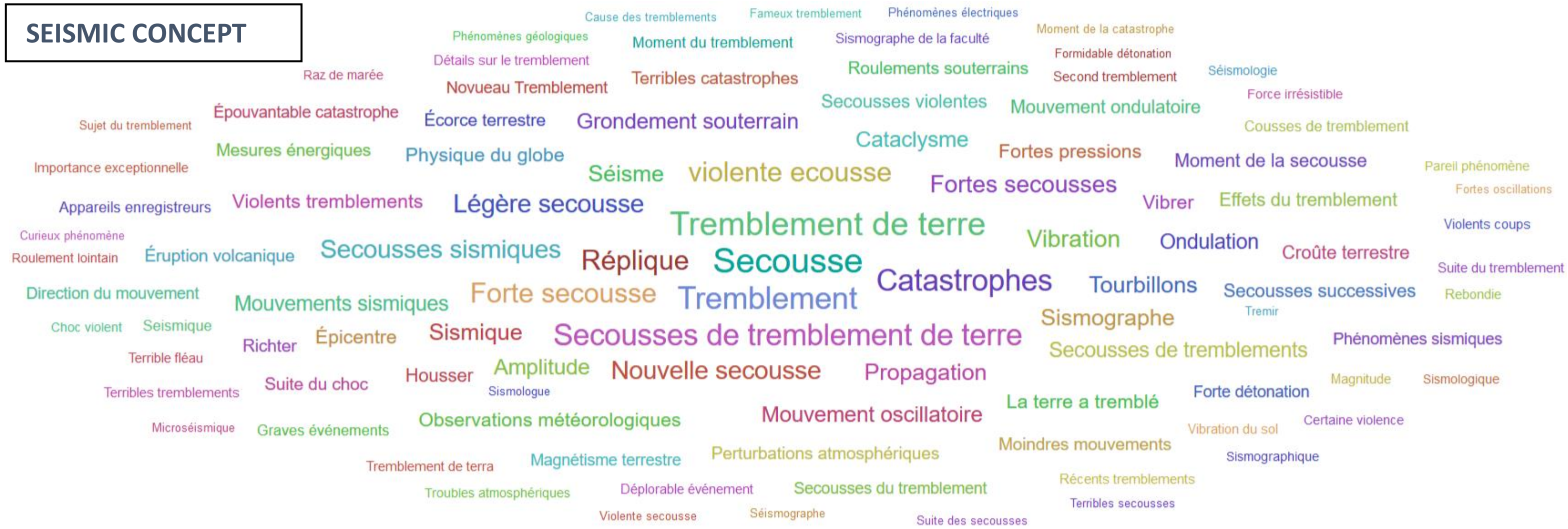
bourdonnement, bruit sourd, canon, cliquetis, détonation, explosion, grondement, sifflement, mugissement, fracas, brait, bruissement, roulement, coup, craquement, tintement, rumeur, trépidation, bourrasque, canonnade, bruit, ronflement

DIVINE CONCEPT

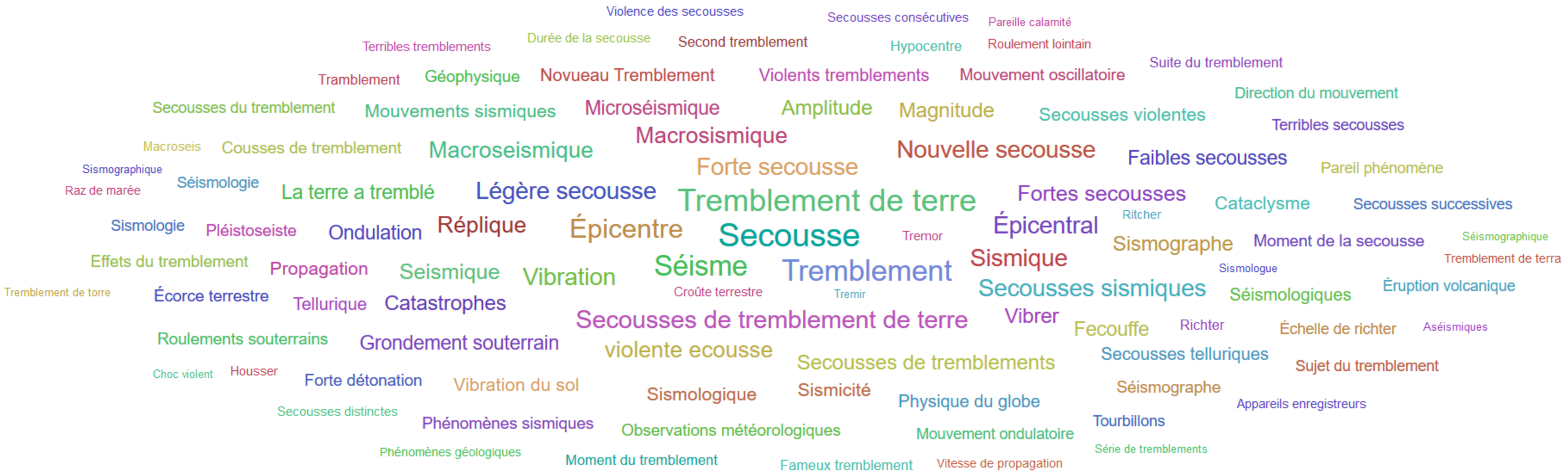
miséricorde, dieu, maléfice, démon, divin bienfaiteur, diluvium, sacrifice, sacrifier, prophétie, jugement dernier

Appendix 2: Frequency of appearance of words from seismological ontology

SEISMIC CONCEPT

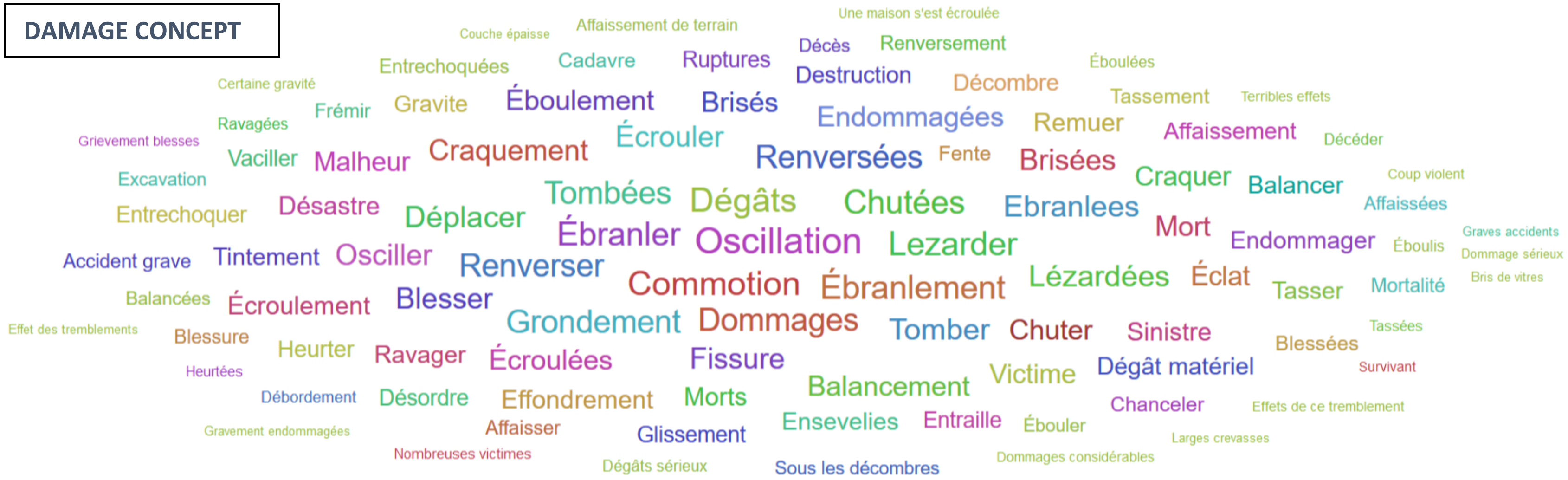


GALLICA

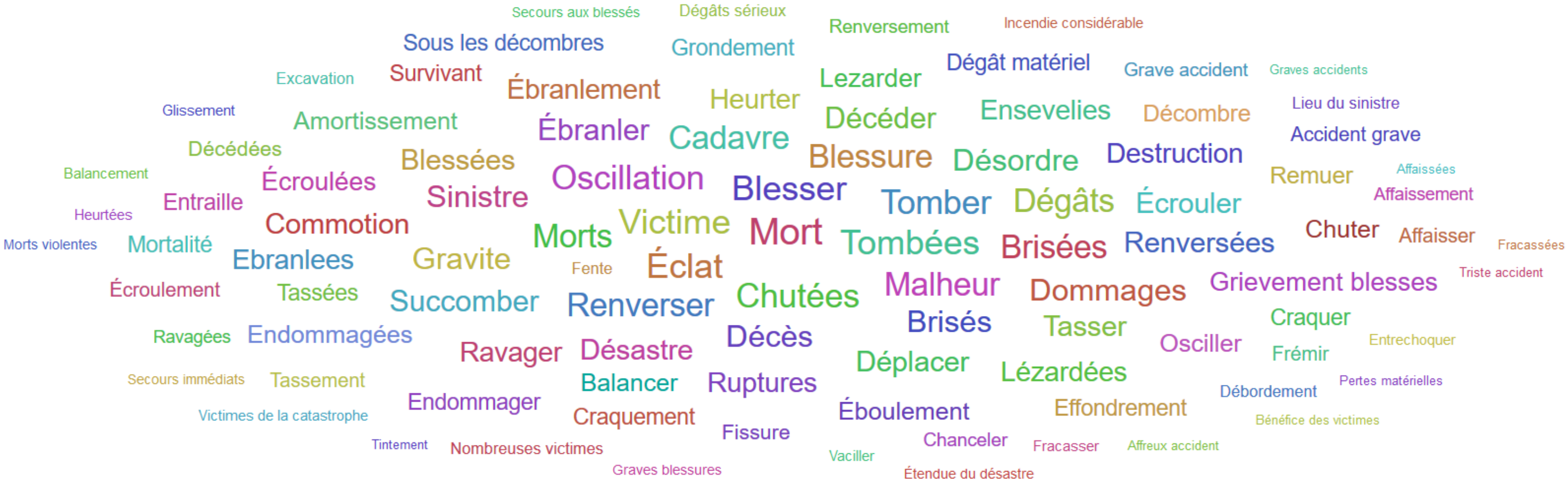


SISFRANCE

DAMAGE CONCEPT

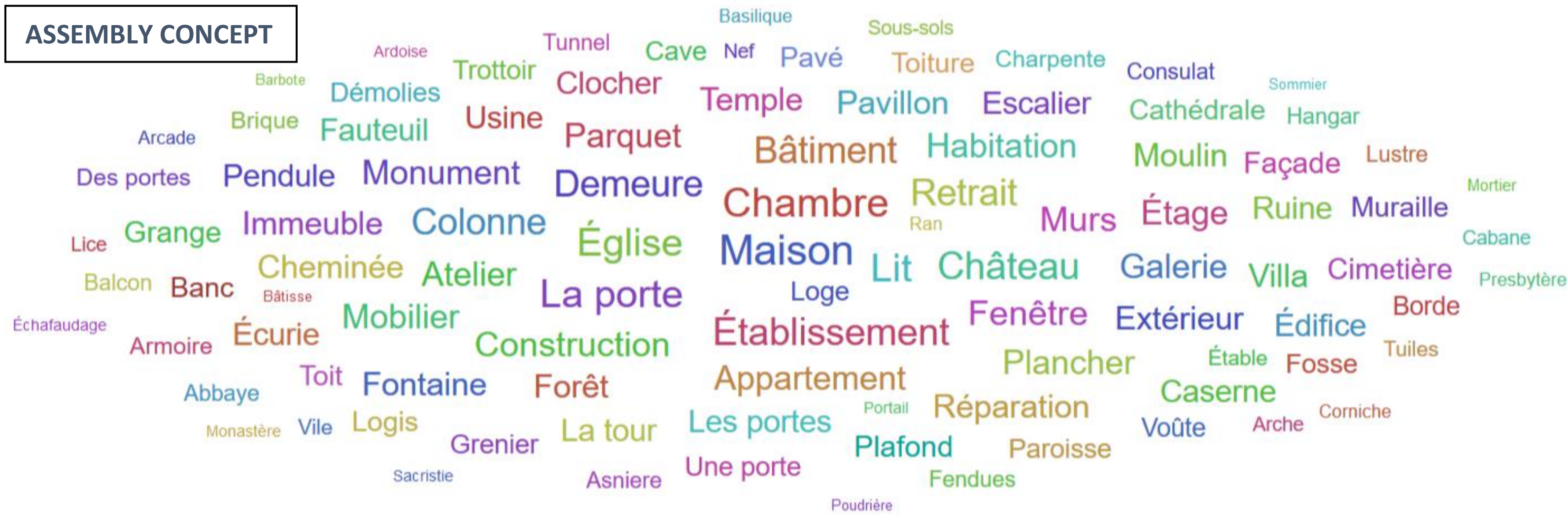


GALLICA

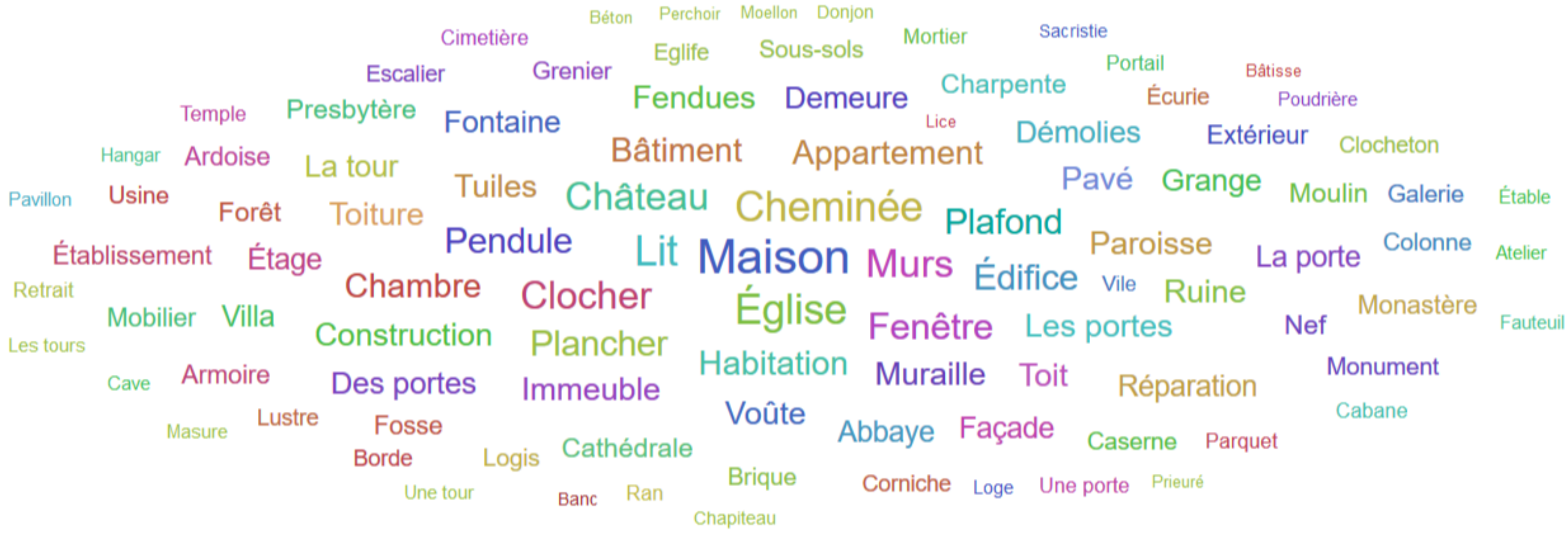


SISFRANCE

ASSEMBLY CONCEPT

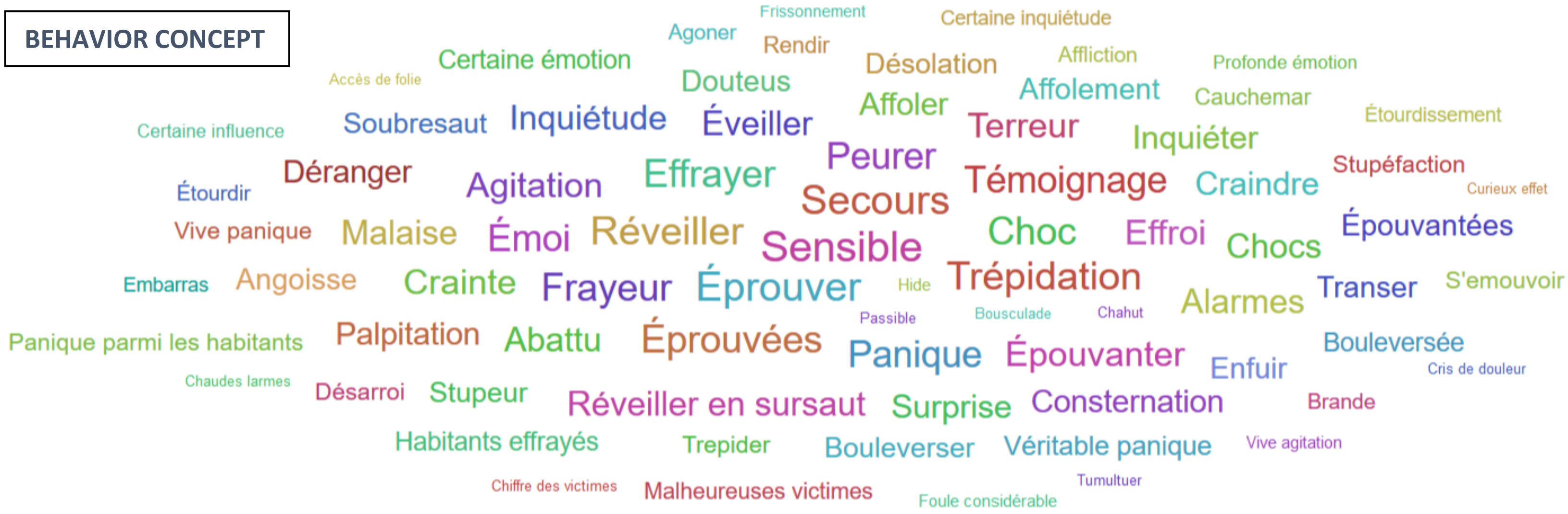


GALLICA

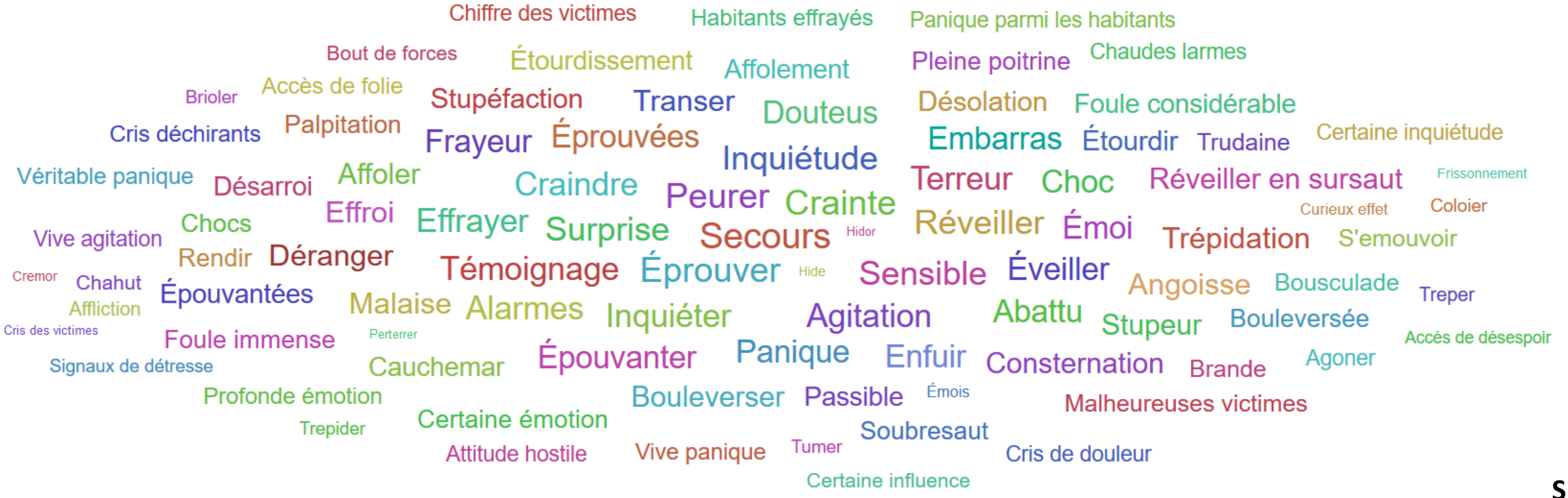


SISFRANCE

BEHAVIOR CONCEPT

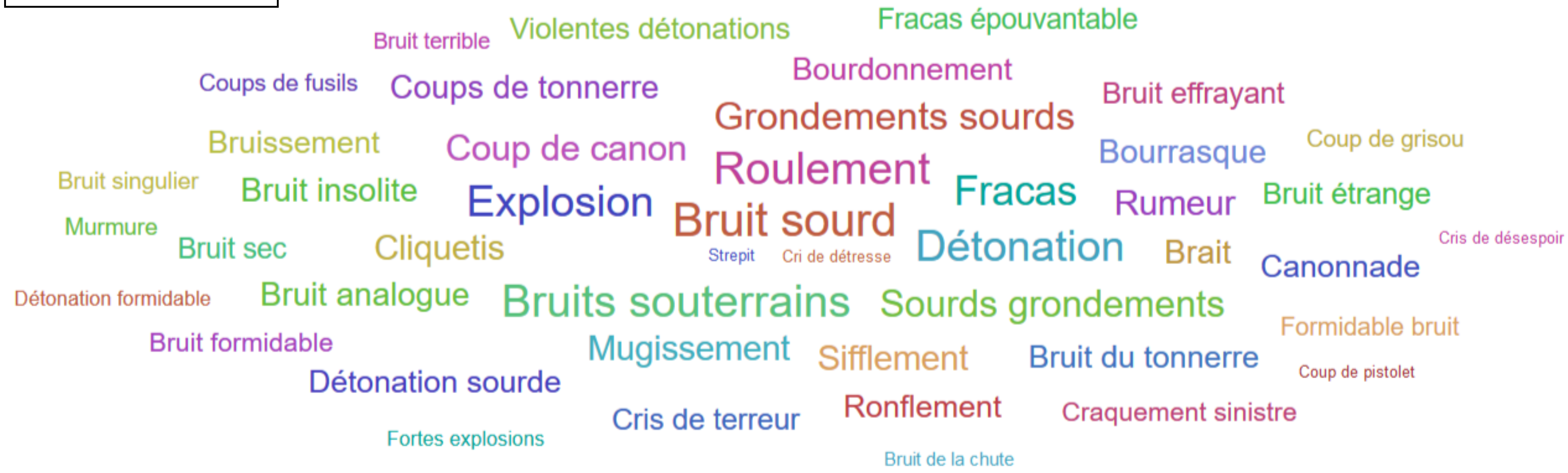


GALLICA

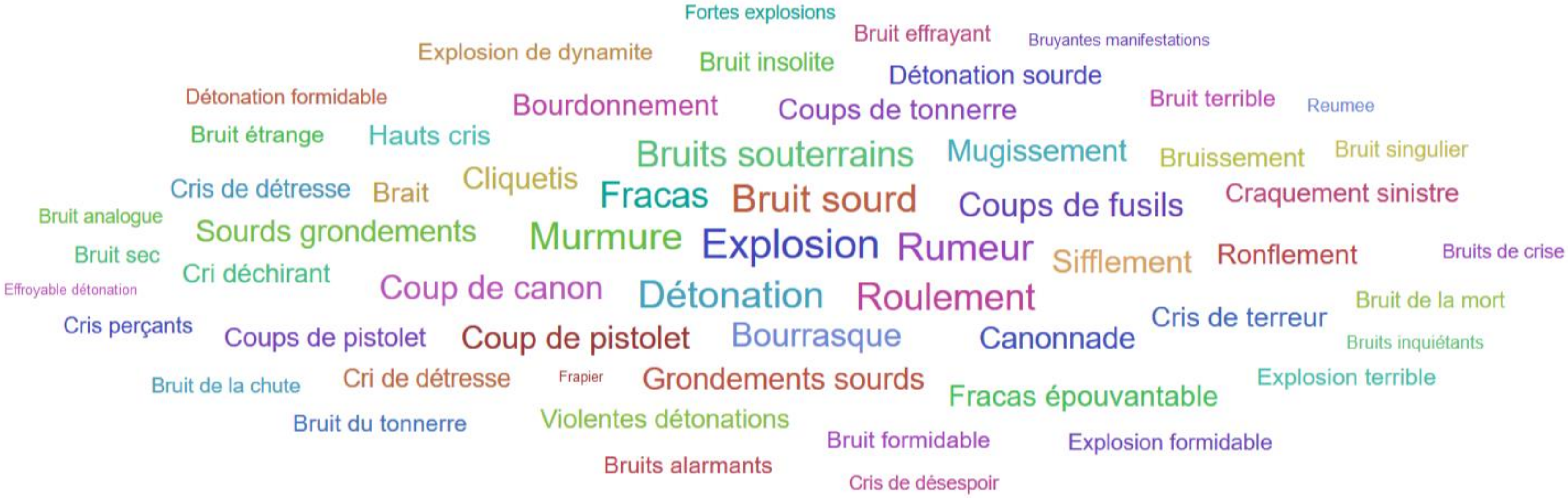


SISFRANCE

NOISE CONCEPT

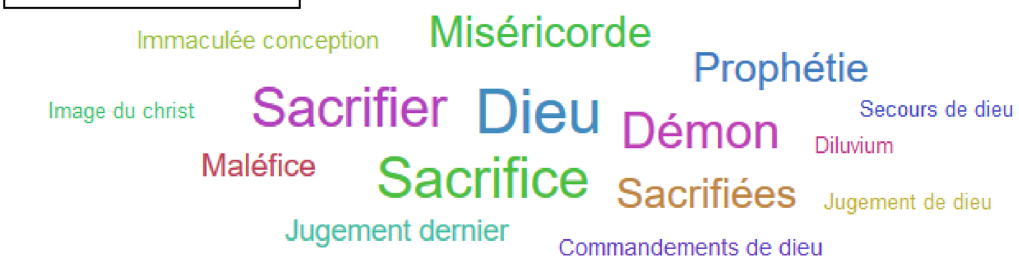


GALLICA

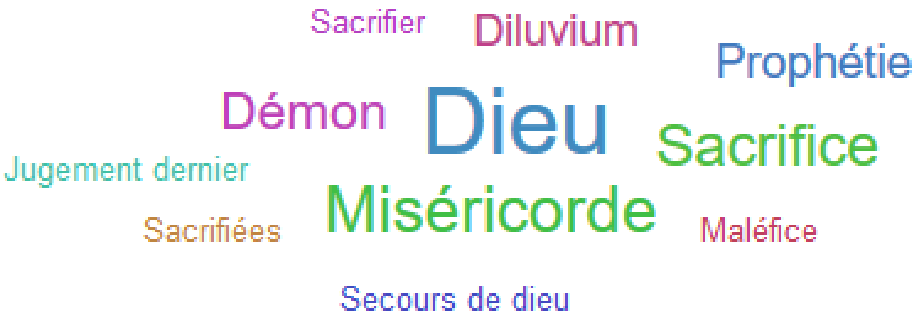


SISFRANCE

DIVINE CONCEPT



GALLICA



SISFRANCE

Appendix 3: List of unknown earthquakes felt in mainland France

GALLICA	DATE	LOCALITIES
Link Document	05/09/1892	Ajaccio
Link Document	??/??/163?	Avignon
Link Document	13?	Avignon
Link Document	09/09/1802	Beauvais
Link Document	09/09/1802	Beauvais
Link Document	02/02/1427	Bordeaux
Link Document	Avant fin du Ivème siècle	Brest
Link Document	30/12/1776	Caen
Link Document	?/03-04/1709	Cigné
Link Document	07/12/1831	La Trinité
Link Document	28-29/12/1863	Lagnes
Link Document	28-29/12/1863	Lagnes
Link Document	25/07/1892	Lens, Arras
Link Document	10/07/1879	Mortain
Link Document	11/11/1891	Nantes
Link Document	05/07/1877	Nantes, Trentemoult, Richebourg, Chapelle-sur-Indre
Link Document	19/01/1405	Narbonne
Link Document	06/08/1580	Nice?
Link Document	05/02/1880	Pau
Link Document	20/10/1852	Poet Laval
Link Document	??/10/8??	Poitiers
Link Document	1467	Saint-Saulve (Montreuil)
Link Document	31/10/1941	Sévignac-Meyracq
Link Document	02/09/1802	Strasbourg
Link Document	??/02/1903	Tarbes
Link Document	22/02/1799	Vannes

Appendix 4: Qualitative results table

New information collected in new documents

NUMEVT	Io	QIEVT	QPOS	
120009	6	B	D	
1	New details about dammage on an already know locality			
2	New details about localisation of the tremor but without a specific location			
610006	6	D	B	
1	New details about the tremor: hour			
2	New details about the tremor: hour			
21	New details about the tremor: hour			
380040	6.5	C	C	
24	New details about feelings on an already know locality			
42	New details about feelings on an already know locality			
1150020	6.5	C	C	
2	New location: La Baule			
8	New location: La Baule			
9	New location with details about feelings and dammage: Besné			
1120046	9	K	C	
1	Structural dammage influence on the future			
500023	6.5	C	C	
5	New location with details about feelings and dammage: Vimoutiers			
840081	7	D	C	
1	New location: Roquemaure. New details about feelings and dammage on an already know locality			
130064	6	C	C	
1	New location: Toulon			
2	New details about feelings and dammage on an already know locality			
130059	6	C	C	
1	New locations with details about feelings and dammage: Toulon, Vernières, Charleval			
2	New details about feelings on an already know locality			
630028	7	B	D	
2	New locations with details about feelings and dammage: Vic-le-Comte, Mozun			
50032	7	A	C	
2	New détails about tremor			
380010	6	D	C	
1	New details about feelings and dammage on an already know localities			
670005	6	C	E	
1	New details about dammage			
1100003	6	B	C	
1	New locations: Maroilles, Fayt, Avesnes, Ohain			
740024	7	C	D	
1	New details about precise localisation and dammage on an already know locality			
130054	7.5	E	E	
1	New details about feelings and dammage on an already know localities			
760040	6	A	C	
1	New location: Saint-Jouin			
630042	6	E	E	
1	New locations with details about feelings and dammage: Neschers, Maringues			
1120006	7	E	E	
3	New location with details about feelings: Hautecourt, Champvert			
1100002	7	B	C	
3	New locations with details about feelings and dammage: Huy, Tirlemont			
690025	6	C	B	
7	New locations: Roman, Valence			
740009	7	A	D	
2	New details about feelings on an already know locality			
1150019	6	C	C	
1	New locations: Damville, Benouville			
2	New locations with details about feelings and dammage: Vésinet, Asnières, Clichy			
8	New locations: Saint-Servan, Paramé			
26	New details about feelings			
1110061	8.5	K	C	
1	New details about dammage			
740035	7	B	C	
6	New location with details about feelings and dammage: Moutiers			
640292	7	B	B	
1	New location with details about feelings and dammage: Arcachon, Soustons			

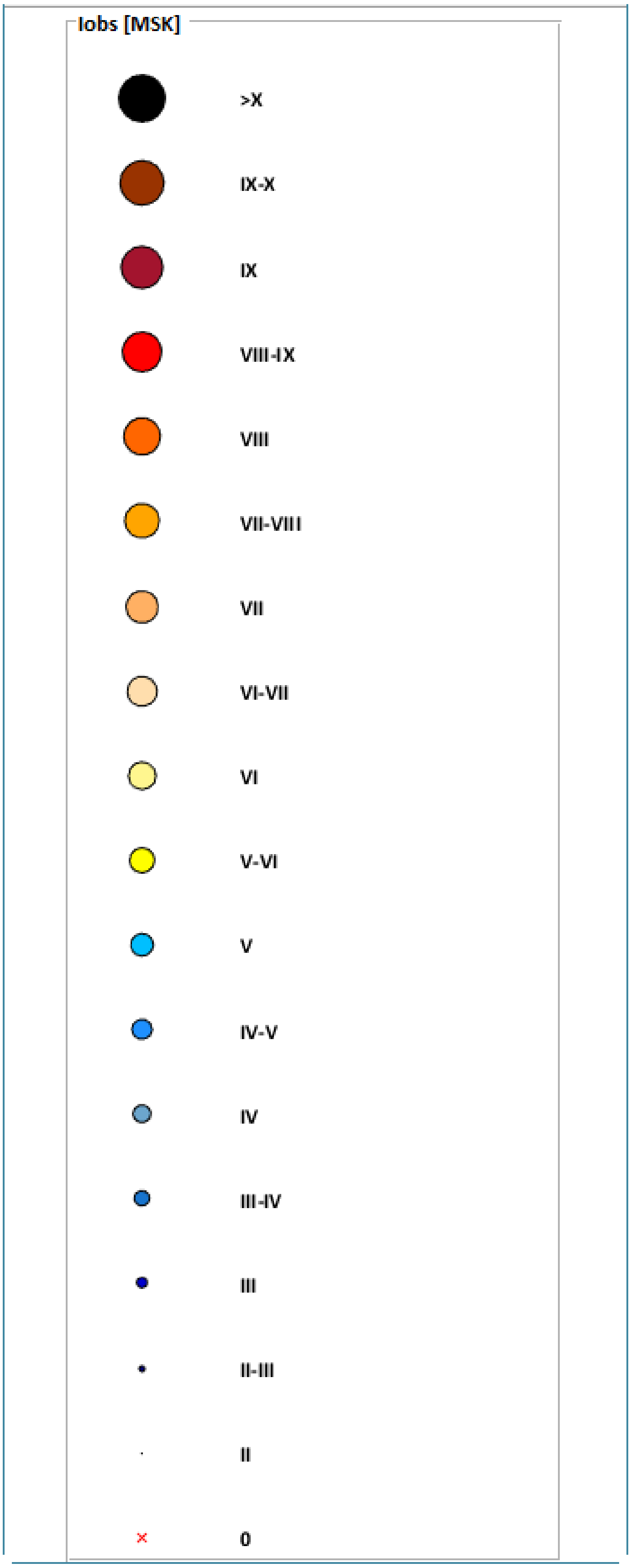
Appendix 5: Creation of observation points (IDPs) and proposed intensity value

Creation of observations in new location			
Numevt	*Location*	*Qiobs*	*Iobs*
1150007	ROWHEDGE (COLCHESTER)	B	7
1150007	WOOLWICH (LONDON)	B	4
1150020	LA BAULE	A	
1150020	BESNE	B	5
610006	CORBEIL-ESSONNES	B	3
840081	ROQUEMORE	A	7
130064	TOULON	C	2.5
130059	CHARLEVAL	A	
130059	VERGNERES	A	
130059	TOULON	B	2.5
630028	VIC-LE-COMTE	A	5
630028	MAUZUN	A	7
1100003	MAROILLES	A	
1100003	FAYT	A	
1100003	AVESNES	A	
1100003	OHAIN	A	2.5
760040	ST-JOUAIN	A	4
630042	NESCHERS	A	5
630042	MARINGUES	A	3
1120006	HAUTECOURT	A	2
1100002	HUY	A	7
1100002	TIRLEMONT	A	7.5
690025	ROMANS	B	
690025	VALENCE	B	
1150019	DAMVILLE	A	
1150019	BENOUVILLE	A	
1150019	VESINET	A	6.5
1150019	ASNIERES	A	2
1150019	SAINT-SERVAN	A	
1150019	PARAME	A	
640292	ARCACHON	A	
640292	SOUSTONS	A	
500023	VIMOUTIERS	B	5
130064	AIX-EN-PROVENCE	A	6
130064	LA ROQUE D'ANTHERON	A	
130064	CHARLEVAL	A	

Appendix 6: Modification of observation points (IDPs) and proposed intensity value

Actualisation of observations in location already known in SisFrance 2017					
Numevt	*Location*	*Qiobs (2017)*	*Iobs (2017)*	*Qiobs*	*Iobs*
840081	AVIGNON	A		A	4.5
130064	LAMBESC	C	6	A	7
130064	SAINT-CANNA	C	6	A	7
130064	ROGNES	C	6	A	7
380010	GRENOBLE	A	5	A	6
670005	STRASBOURG	B	6	C	6.5
1110061	STRASBOURG	A	5	A	6
740035	MOUTIERS	B	5	A	4

Appendix 7: Intensity scale (IDP)


Iobs [MSK]
>X
IX-X
IX
VIII-IX
VIII
VII-VIII
VII
VI-VII
VI
V-VI
V
IV-V
IV
III-IV
III
II-III
II
0