IRSC 2022

INTERNATIONAL RAILWAY SAFETY COUNCIL

SEVILLA, OCTOBER 16-21, 2022

Machine learning
applied to Event report archives to improve Safety

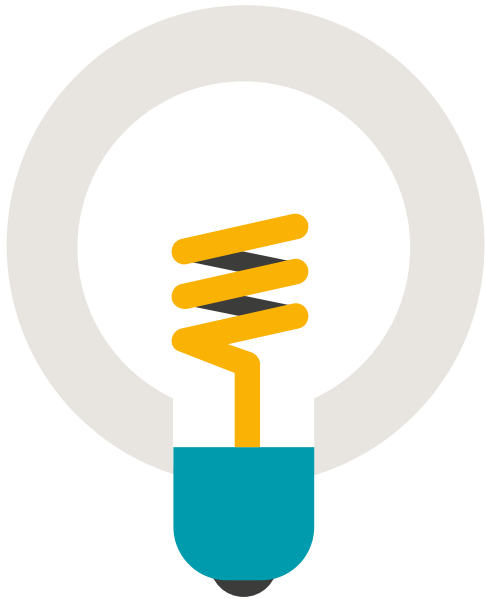**GUILLAUME FOEILLET**   SNCF – Safety Department

on behalf of the 'SMART STUDIO' – **SNCF RESEAU DG_NUM**

- SMART studio's aim is to improve the railway knowhow

taking advantage of new technologies. Safety is on top of the file.

The 'Smart Studio' uses New technologies for data analysis to contribute to bring the digital features to the whole railway domain knowhow :

+ Smart Language : Automated **P**rocessing of text data in **N**atural **L**anguage (NLP)

+ Smart Maintenance : Collection and data analysis to feed predictive maintenance software tools

+ Smart Certification : Information Storage and Exchange in a secure way (Blockchain)

10 Internal customers

6 On-going projects

IRSC Spain

EUROPEAN UNION AGENCY FOR RAILWAYS

AESF AGENCIA ESTATAL DE SEGURIDAD FERROVIARIA

adif

SNCF

CONTENTS

SNCF explains here how the records from the past safety events can now be efficiently used to improve the analysis and make appropriate investment decisions for safety, with all the required justifications - with a new tool using a semantical approach and Machine learning principles.

CHAPTER 1
Application context and identified needs

APPLICATION CONTEXT AND IDENTIFIED NEEDS

In-house archives have been initially created to address an 'administrative' traceability requirement to record the events and all useful details.

As long as the record is filled in, there is legal compliance, even if poorly filled-in. Free text zone is proposed as incentive to build record.

Records are useful and in-house tools are efficient to retrieve all information relative to /one precise event /with matching key words.

Main application is to retrieve data for judiciary purpose, or to perform in-house detailed analysis on root causes on specific event.

Tool's HMI is less efficient when a transverse need requires to make sure to parse all relevant records which may contain different key words.

Difficult to establish statistics and cost-efficiency ratios over complete archives, for instance to prepare and justify safety investments.


Issue also comes from the fact that records contain free text (unstructured data), while Tools only propose filtering with predefined set of words.

Records are filled-in with various formats, texts, scans of handwriting, pictures, diagrams – filters are not enough powerful to parse in them all.

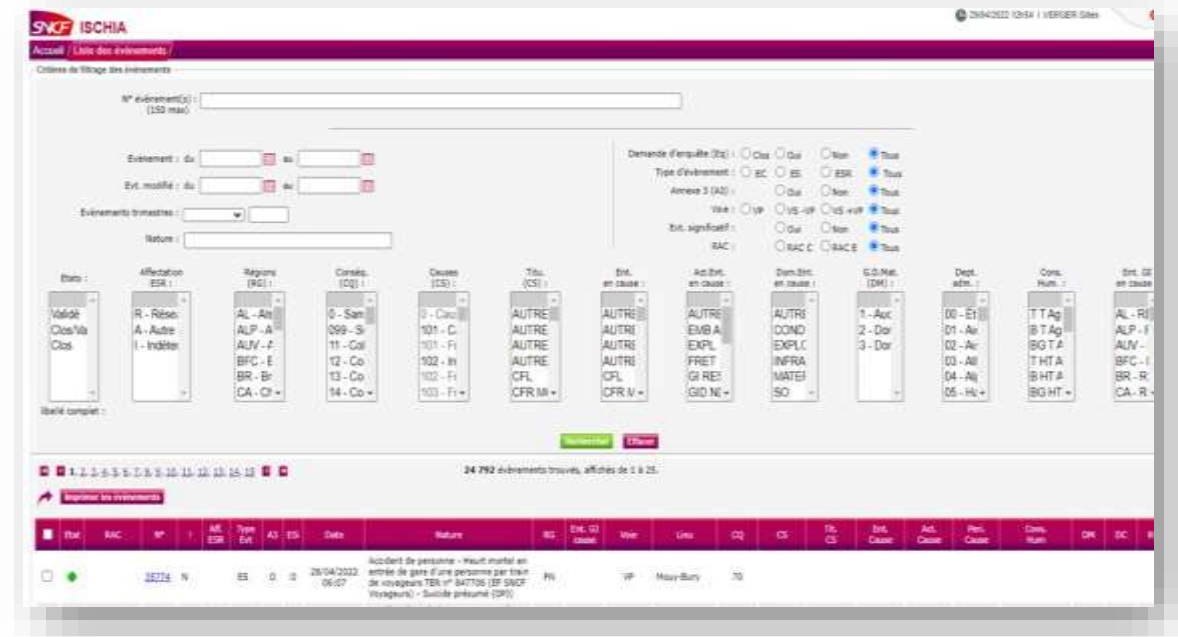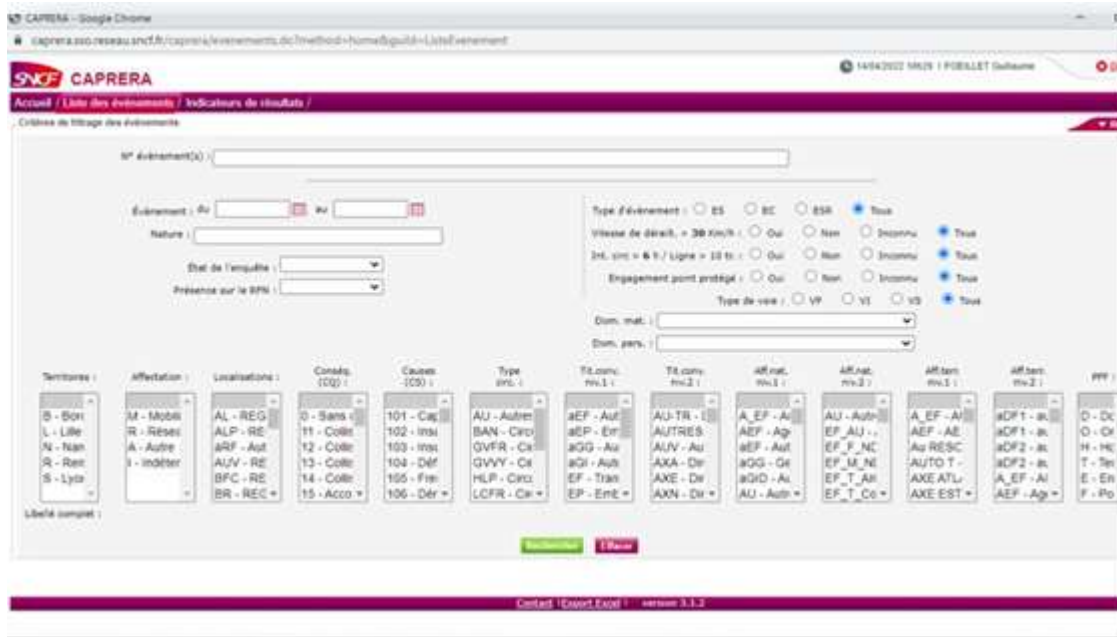We are all now familiar with browser's use in our everyday life…

We are now frustrated not having the same feature for parsing our safety events!
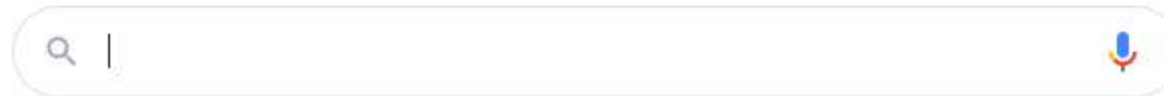
SNCF SAFETY EVENT DATABASES : ISCHIA & CAPRERA

-More than 10 years of railway safety events for

[SNCF RESEAU] - Infrastructure Manager side &

[SNCF VOYAGEURS] - Pssgr Train Operator side

- Currently, information can only be accessed via limited predefined filtering criteria or via the safety event tag number

- Free-chosen key words can not be provided as filtering seed

➔ **Stakes** : optimise the use of data, switching from a customised research mode to a multi-purpose / multi-criteria request mode

(is it possible to...) IMPROVE SAFETY EVENT MANAGEMENT – with archives?

Yes, it's possible to make <u>much better use of the events from the past</u>

to fully understand and better invest / anticipate / avoid Railway Safety Events

« THAT'S THE WAY IT WAS, STILL IS » ...

§ A rather **difficult** access to the existing databases

§ **Expensive / sweating enquiries,** they are time-consuming for their process and need ressources

§ Seldom or no true **use** once the reports are published

during and following the enquiry period.

§ **No real internal benefits for the knowhow improvement**

and no detection or deduction of local / regional exceptions

NO...THERE ARE « USEFUL STEPS » TO PROGRESS THE MATTER

- Reinforce the **access** then the **quality** of the information databases

- Perform more targeted analysis from one or several

  given historical contexts to progress on root causes and potential barriers

- Provide assistance to **decision making** by stakeholders (investments)

- Guarantee **efficiency** and post-incident **management**

IMPROVE SAFETY EVENT MANAGEMENT – with archives thanks to MACHINE LEARNING!

To achieve this, SNCF has decided to explore and take advantage from Natural Language Processing (NLP)

It takes advantage of all the techniques which use **computers** to analyse data, determine the **semantical** similarity between words and translate between languages. It applies to **written languages** but also to **speech**.

It can lead to Machine Learning or automated learning, a « concept which makes a **machine** able to learn from its own experiences »

Semantical Analysis consists in determining the exact meaning of words or a set of words by a detailed analysis of the **combination of words and the context,** it can help machine to consider by itself all other words with similar meaning.

- Usual application contexts of Machine Learning and Semantical analysis:
> Research programs / Text contraction / Automated translation, Statistics, Analysis & Decisional tools

- This domain is associated to concepts such as: 'Corpus' of data / 'Thesaurus' / 'Lexicon' / 'Labels'

IMPROVE SAFETY EVENT MANAGEMENT – launch of a proof of concept (POC)

01 | **Data mining aiming at all useful archives**

02 | **Analysis and Processing of extracted data**

03 | **Production of forecasts and interpretation**

> *Collect useful data to obtain from tool expected deliverables*

> Semantical analysis

> Machine Learning
> Data visualisation

Technical approach - 3 successive Work Packages

TWO NEW FUNCTIONALITIES

> **Identified target Users :** Authorised employee has standard access to data, either standardised or displayed as if standardised.

> Obtain **Automated filled-in safety reports** (standardisation)
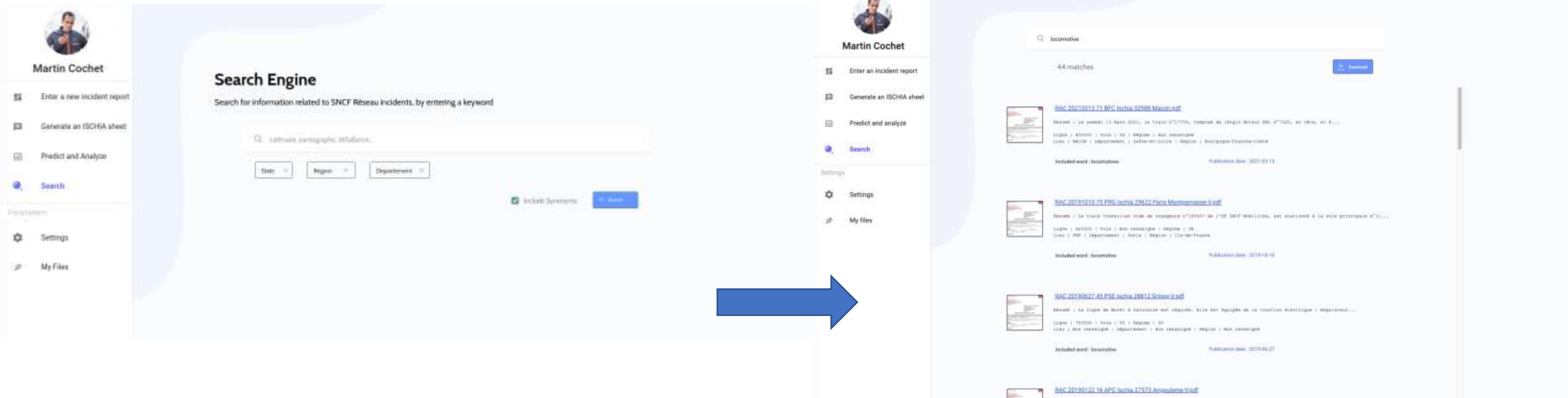> Make good use of a **Smart search engine**

A RELIABLE STATISTICAL TOOL

> Authorised employee provides as input some contextual elements and receives as output the **potential root causes** of accident, as well as their weight in archives

> **Identified target Users** : eg. Regional managers willing to decide and prioritise maintenance investments.

A DECISION ASSISTANCE TOOL

> *After the safety event*: Identification of **low noise signals** and assistance in solving long time pending enquiries

> *In real time* : Authorised employee provides as input contextual elements and receives as output the **potential root causes** of accident.

> **Identified target Users** : eg. Local Safety Managers, Enquiry managers, Overhead line event specialists, entities at all level within Infrastruture Manager.

IRSC Spain

EUROPEAN UNION AGENCY FOR RAILWAYS

AESF AGENCIA ESTATAL DE SEGURIDAD FERROVIARIA

adif

SNCF

Design a smart search engine to ease searches   in order to…

To feed the Proof-of-Concept (POC)   ns demonstrate efficiency of the proposed strategy

- More than 1300 documents have been extracted from Databases for testing purposes

- **Functionalities:** Search is launched with one or more key words / Synonym analysis / additional filtering possible

- **Evolutive releases** (filters, results display, results downloading…)

[Non-structured data] are now also taken into account

- Safety event context (circumstances)
- Root causes
- Consequences
- Risks
- Risk Control Measures (**RCM**) decided and set

**It offers some prediction opportunities**

- type of event (precise set of root causes)
- Identification of cumulative consequences
- Risk control measures to prevent or mitigate root cause
- RCM to prevent or mitigate consequence
- Type of rolling stock with the type of cause / of consequence
- Location of incident (regional, local) with root cause
- Impact of safety event given the root cause
- Type of signal passed at danger given the root cause
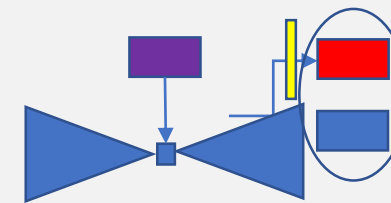
Find out [cause & cause]
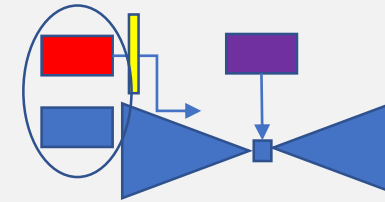
Improve existing Bow-ties
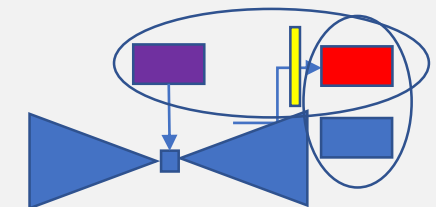
Find out [cause & safety event]
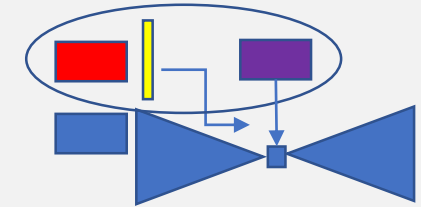
Improve existing Bow-ties
Create new Bow-ties

Find out [consequence & consequence]

Improve existing Bow-ties

Find out [safety event & consequence]

Improve existing Bow-ties
Create new Bow-ties

...→ help at all stages with this [search / statistical / decision-making] assistance tool

ENQUIRY MANAGER		LOCAL SAFETY MANAGER		TOP LEVEL SAFETY MANAGERS
①	②	③	④	⑤

AUTOMATED SAFETY EVENT CHECK-LIST FILL-IN

ASSISTANCE TO THE IMMEDIATE SAFETY REPORT

PRELIMINARY FILL IN OF THE EVENT RECORD

ASSISTANCE TO THE FINAL SAFETY REPORT

INVESTMENTS

EASE THE COLLECTION OF ALL DATA

Provide thanks to the tool all contextual information to describe the safety event
- semi-automated fill-in of the check-list immediately after the safety event

MAKE GOOD USE OF THE DATA FROM THE PAST

Look for former situations similar event in order to :
+ Suggest some protective actions
+ Identifiy who should be called

EASE THE PARSING

Based upon the Immediate safety report

+ automated fill-in of the event description and protective actions set

ASSISTANCE TO SAFETY DOCUMENTS WRITING

+ Wrap-up of details noticed / conclusions approved by local safety manager
+ Prédiction of the root causes of event on the basis of what has been collected in the past
+ Preliminary fill-in of final report document aligned with company template

DECISIONS

+ Justified and documented investments for safety

IRSC Spain

IRSC 2022
INTERNATIONAL RAILWAY SAFETY COUNCIL
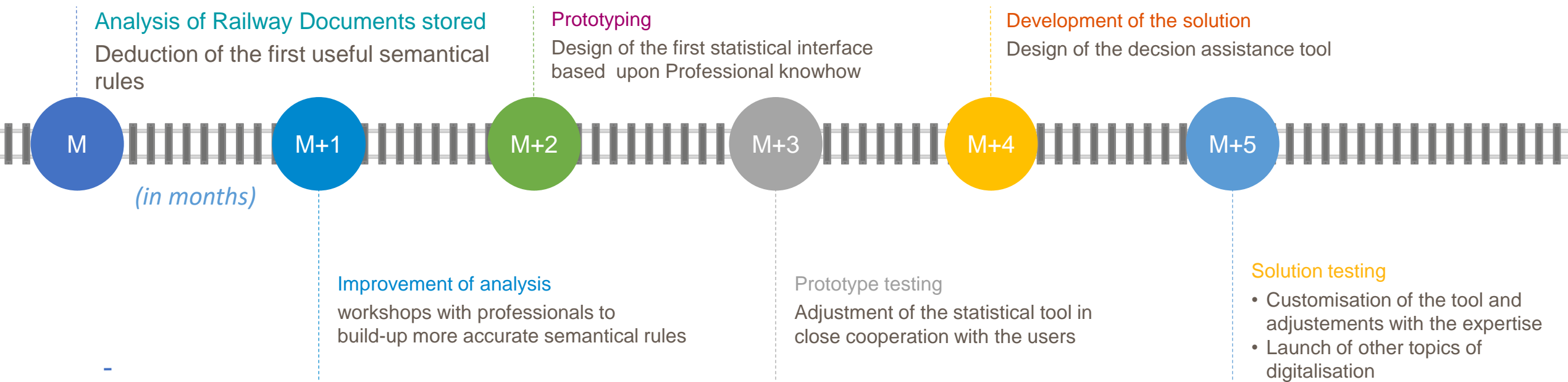**SEVILLA, OCTOBER 16-21, 2022**

EUROPEAN UNION AGENCY FOR RAILWAYS   AESF AGENCIA ESTATAL DE SEGURIDAD FERROVIARIA   adif

SNCF

CHAPTER 2

RESULTS OF THE PROOF OF CONCEPT (PROPOSED SOLUTION)

ROADMAP of the Proof Of Concept

Analysis of Railway Documents stored
Deduction of the first useful semantical rules

Prototyping
Design of the first statistical interface based upon Professional knowhow

Development of the solution
Design of the decsion assistance tool

M   M+1   M+2   M+3   M+4   M+5

(in months)

Improvement of analysis
workshops with professionals to build-up more accurate semantical rules

Prototype testing
Adjustment of the statistical tool in close cooperation with the users

Solution testing
- Customisation of the tool and adjustements with the expertise
- Launch of other topics of digitalisation

-

REQUIREMENT free access to (the usually restricted access) data stored in safety databases (obtained)

REQUIREMENT budgets for a more robust design of the tool applied to all data, derived from the POC (obtained)

Both did require some internal lobbying to convince high level management and official approval (➜ non-technical aspects)

IRSC Spain

EUROPEAN UNION AGENCY FOR RAILWAYS

AESF AGENCIA ESTATAL DE SEGURIDAD FERROVIARIA

adif

SNCF

WORK PERFORMED DURING POC

Step 1 : **Exploring data and processing the safety reports**

There is a stong need to analyse a set of reports to extract:

- **Structured text zones**:
  Date, location, Line number, Track, Train type, …
- **Free text zones** : story telling, context, consequences, risks, risk control measures ...

Step 2 : **Setting rules of semantical analysis**

- Definition and construction of a **Thesaurus**
- Identification of **key words**, relevance ranking level

Step 3 : **Include requirements for the Python script**

- Documents **pseudo-cleaning** and data **import**
- Language multiple feature analysis
- Transfer of all useful (railway) concepts

Step 4 : **Perform iterations for AI and Machine Learning**

- Automatic classification
  (monitored learning, multiple tag classification)
- Statistics and Predictions relative to each event

Step 5 : **Data display**

- Display of data to reveal difference and feed decision
- Delivery of automatic tailored predictions

> **Note**
>
> - Experts should to be involved as soon as step 2 in order to obtain validation of the semantical choices
> - Afterwards, the items that could be predicted have to be defined together with the end customer, sponsor of the project.

BEST EFFORT WITH MINIMAL DATA

> **! Note**
>
> During the POC:, it has not been planned to:
>
> - Process the <u>complete</u> set of files (only the amount required to demonstrate the feasibility and the quality of the process)
> - Set up a thesaurus as large as possible to address all possible topics of analysis.

It has however been absolutely necessary to optimise the analysis of all the available data once the newly extracted data were copied into a distinct data field. The results of the POC did depend on the amount of data extracted from database on which the tool will be applied later on...

This analysis has enabled to :

- Collect a reliable well structured and non-structured information out from the any kind of files available

- Set up a mechanism to tag with index the extracted content to initialise and feed a Thesaurus, customised for the types of expected analysis

- Demonstrate feasibility of enhancing each safety event file with additional information thanks to the thesaurus (and file classification)

- Demonstrate feasibility of performing multicriteria investigations from existing safety event files, taking advantage from information obtained by algorithm and also parsing through free text

- process data with selection of analysis topics, to be defined given the set of data to be used for the analysis.

Extract from typical Root cause analysis reports.
Chartered ('structured') document and also any kind of ('as is', 'unstructured') files, attached to it

RAPPORT D'ANALYSE CAUSALE n°11450

Avertissement : Le présent rapport est réalisé dans le cadre du décret 2006-1279 du 19/10/2006. Ce rapport décrit les circonstances, les conséquences et les causes directes et indirectes d'un évènement de sécurité. Il ne vise pas à déterminer des responsabilités à la suite de cet évènement.

Identifiant de l'événement : 27-09-2012 / 06h00 / 905000 / Collision du train TER SNCF n°885652 avec un rocher suivi du déraillement du 2ième essieu du 1er bogie du train (sans blessé).

1. DESCRIPTIF

N° fiche ISCHIA : **11450**   Lieu : Voie Unique Veynes-Grenoble Pk 229,950   Voie : VP
Régime d'exploitation : VU   Cantonnement : CAPI
Train : 885652   Activité : Voyages   EF : SNCF
Le train est composé d'un élément ATER X73636.

2. CIRCONSTANCES

Le Jeudi 27 Septembre, le train TER 885652 de l'EF Voyages SNCF circule normalement sur la voie unique entre Veynes- Dévoluy (origine départ 5h50) et Grenoble (arrivée 7h43).

Le train 885652, après avoir desservi la gare de Aspres-sur-Buëch (5h55/5h56), reprend sa marche à la vitesse limite de 80 km/h. Vers 6h00, en sortie de courbe, le conducteur du train 885652 aperçoit au dernier moment un bloc de rocher important sur la voie de circulation. Le conducteur freine d'urgence, mais ne peut éviter le heurt au Pk 229,950 à la vitesse de 75 km/h. Sous le choc, le deuxième essieu du 1er bogie déraille et le train, sans se coucher, s'immobilise au Pk 229,850.

Le conducteur avise le SGTC de la collision et du déraillement. Aucun blessé parmi les voyageurs et le personnel à bord.

3. CONSEQUENCES

Gravité des dommages :   Personnes transportées : RAS

DIRECTION DE LA CIRCULATION FERROVIAIRE   SNCF

7. Conclusions du GID

Rapport conclusif : le 27/09/2012, un déraillement s'est produit au Km 229.950 suite à heurt de bloc par le train n°885652 vers 6h00. Les passagers ne sont pas blessés et sont transbordés par taxis. Le train a endommagé la voie sur environ 100 mètres.

Le site classé ouvrage en terre sensible se situe entre les Km 229,100 et 229,950, avec notamment la présence d'écran en traverses en bois au Pk 229,100 et 229,750.

Un rocher (masse estimée à 800 l) s'est décroché de la ligne de crête de la tranchée rocheuse, a parcouru 300 mètres et a fini sa course sur la voie ferroviaire (hors écran). La cicatrice de chute du bloc ayant atterri en voie se trouve en partie supérieure de versant et donc hors emprise ferroviaire.

La visite réalisée le 1er octobre 2012 par des représentants de l'Infra V a permis de mettre en avant les points suivants :

- de confirmer qu'il n'y avait pas de risque de chute de blocs imminente pour des blocs de volume conséquent (environ 1 m3 ou supérieur) au niveau de la cicatrice de chute. Il a donc été considéré que la situation initiale (avant incident) était rétablie. Sur la base des conclusions, la vitesse de circulation a été relevée de 10 à 40 km/h ; le ralentissement à 40 km/h étant nécessaire dans l'attente de la fin des travaux de voie. Ces derniers seront effectués pendant la fermeture de ligne de novembre 2012.

- le tiers propriétaire des fonds dominants devra être identifié et prévenu de l'incident et de ses conséquences. Il devra également être prévenu qu'en cas de dommages futurs, sa responsabilité pourra de nouveau être recherchée.

Detail view of the extracted structured fields

Structured data extraction

• :

- • **Document number**
- **Date of safety event**
- **Where event took place :**
   region, department,
     kilometric point
- **Train operation mode**
- **Train number**
- **Track type**
- **Type of power car/engine**
- **Use**
- **Line number / Kilometric point**

Detail view of the extracted non-structured fields

- **Acquisition of non-structured data**

  - ☐ **Safety event context (circumstances)**
  - ☐ **Causes, root causes**
  - ☐ **Consequences**
  - ☐ **Type of risk(s)**
  - ☐ **Decisions made, actions to be taken**

Dont blessés graves	0	0	0	0	0	0	0	0

Version du : 21/01/2020   Page 1 sur 2

Information immédiate

Description de l'évènement :

Contexte :
Le train n° 868379 de L'EF SNCF Voyageurs vient d'effectuer une mission entre Objat (19) et Brive (19). Il es parti à l'heure d'Objat (19h54) et arrivé avec 1 minute de retard à Brive (20h24).
Ce train est reçu voie E, en gare de Brive.
La voie utilisée pour effectuer la rentrée des matériels au dépôt, est la voie 2D.
La voie 1D est utilisée comme sas de man?uvre, ainsi que pour effectuer des sorties du dépôt.
Les faits :
Une fois le service commercial du train effectué en gare de Brive, le conducteur change de cabine de conduite afin d'assurer sa rentrée au dépôt. Après avoir terminé la préparation, il en avise l'agent de man?uvre présent sur le quai qui retransmet par radio à l'Agent Circulation (AC) de Brive Poste 1.
L'AC trace son itinéraire, mais se trompe en frappant son dialogue. Il commande 214-207 au lieu de 214-205, ce qui a pour conséquence de diriger le train vers la voie 1D au lieu de la voie 2D.
L'autorisation de mise en mouvement est donnée au conducteur.
Le conducteur déclare qu'il s'avance, radio éteinte, vers le signal n° 214, mais n'observe pas correctement la signalisation présentée. Le signal n° 214, lui présente un feu blanc clignotant, alors qu'il aurait dû lui présenter un feu blanc fixe, accompagné de la pancarte "D".
Dans le même temps, l'AC se rend compte de son erreur et essaye de détruire son itinéraire, mais sa commande n'aboutie pas. Il essaye également d'entrer en contact avec le conducteur, à l'aide de la liaison sol-train, mais n'y parvient pas.
Le conducteur du train n° 868379, ayant franchi le signal n° 214 ouvert, poursuit sa marche vers le dépôt, quand il aperçoit devant lui le signal Cv 498,9 fermé. Il freine d'urgence, mais ne peut éviter son franchissement. Il stoppe son train au km 497+700 et avise immédiatement le SGC, ainsi que sa hiérarchie. La situation est figée jusqu'à l'arrivée des dirigeants d'astreinte.
Actions urgentes mises en oeuvre par les exploitants :

SNCF Mobilités :
Conducteur du train n° 868379
- Arrêt d'urgence.
- Avis immédiat au SGC et au Cadre Traction
- Participation à l'enquête.

Astreinte traction :
- Prise immédiate de mesures pour stopper les sorties dépôts.
- Constat de la situation sur le terrain.
- Entretien "à chaud" avec conducteur.
- Relève de l'enregistrement ATESS.
- Participation à l'enquête.
Mesures conservatoires prises par le GI :

Service chargé de la Gestion de la Circulation (SGC) :
- Tentative infructueuse de joindre le conducteur.
- Figeage de la situation.
- Appel immédiat aux astreintes intéressées.
- Désignation d'un dirigeant d'enquête.
- Etablissement d'un Rapport de Constatations Immédiates (RCI).
Autres commentaires / mesures autres mises en oeuvre :

Renseignements complémentaires :
- Le signal n° 214 (signal d'extrémité de la voie E), est un signal muni d'un ?illeton et d'une plaque d'identification "Nf". Il est également muni d'une balise KVB et d'un crocodile talonnable. Lorsqu'il est ouvert, depuis la voie E, en direction de la voie 2D, il présente un feu blanc, ainsi qu'une pancarte "D".
- Le signal Cv 498,9 est un signal de type bas. Sa position constante est la fermeture.

Pièces jointes

Titre	Date de création	Date de mise à jour	Utilisateur	Conf. DSSR
Schéma synthétique Brive la Gaillarde.docx	20/01/2020	20/01/2020	DEROUAU PHILIPPE	N
Plan de situaton.docx	20/01/2020	20/01/2020	DEROUAU PHILIPPE	N
Precisions signalisation.docx	20/01/2020	20/01/2020	DEROUAU PHILIPPE	N

Indexation - Classification

■ Extracted content is tagged with index

- All the words contained in the extracted data are used to build up a Thesaurus specific to the company / railway activity.
- For each meaningful word, one string or more is associated to it, this enables obtaining efficient results with parsing algorithms « by analogies / synonyms ».

- The complete set of words is extracted and used to build a thesaurus for one or two topics to be chosen.
- Any type of data such as weather data, objects found on track, radio communications related to the safety event

! Note

- This steps requires the contribution of experts to help add things that would be wrong or miss then provide validation of the initialised Thesaurus.

- The thesaurus is « blindly » built, which means not having access to the source PDF document, forbidding access to the whole context of a given incident, which could introduce some bias in the content of the thesaurus which would artificially the performance of the processing.

■ Files classification and multiple criteria research

For the files classification,

- The whole set of files which have to be considered for the study, is classified according to the various categories defined in the thesaurus (thanks to analogies / synonyms).

- The design team performs a first analysis to check the consistency of the resultts obtained. Then, an expert is required to check the results and delivers final validation.

For the multiple criteria research, a set of requests is defined to be submitted to the enhanced data base.

These requests will use:

- The structured data extracted for the purpose of the study;
- Classification categories appropriate for research;
- Key words to be found (including miss spelling, but without invoking the set of synonyms).

These requests are partly built « blindly », without any access to the content of the existing files.

Additional work to refine the predictions and statistics

From the extracted data, tagged with index then classified, the algorithm produces :

- **Statistics** : by cross comparing or numbering the amount of identified structured data
- **Predictions** : by cross comparing identified non structured data, taking advantage of the preliminary defined thesaurus.

Examples of predictions opportunities / expert required to provide validation

Types of relevant safety events / root cause

- Identification of cumulative consequences

- Cross comparing actions taken relative to the safety event cause

- Cross comparing actions taken relative to the consequences

- Cross comparing rolling stock type with the root cause(s) / the consequences

- Cross comparing location of safety event (regional, local) with the root cause

- Cross comparing impact of the safety event with the root cause

- Cross comparing the type of signalling with the root cause

CHAPTER 3
Conclusion

The proposed tool will improve the access and the quality of our safety data

This tool is applicable whatever may be the format of safety report archives, database or flatfile. *It could be your tool too...*
It only requires some configuration following the context analysis

It provides to railway domain analysts an easier access to the safety data without modifying them, with two dedicated efficient parts of the tool :

- a **powerful search engine**, similar to a browser, to parse and filter efficiently the existing safety event reports

- an **automated process** to fully implement the new safety event reports, to obtain complete and standardised (structured) safety reports, while reducing the work load.

Obtain figures, check safety facts, find new ones

The user of tool provides as input in natural language all the useful and known details he wants, can refine the and obtains in return all the available data, and also links between root causes, safety event, damage consequences.

Various benefits can be obtained, for top and middle managers :

- Explore data to get hints, good ideas, relevant assumptions

- Dig in the data collected from the field to confirm the assumptions

- Update existing bow-tie models, suggest new bow-ties deduced from true occurrence on the field

- Establish cost/benefit ratios to compare solutions

- Define and justify priorities for safety investments, and deployment of appropriate risk control measures.

IRSC Spain

SNCF

Example of
compared analysis obtained

established with a sample of 150 files
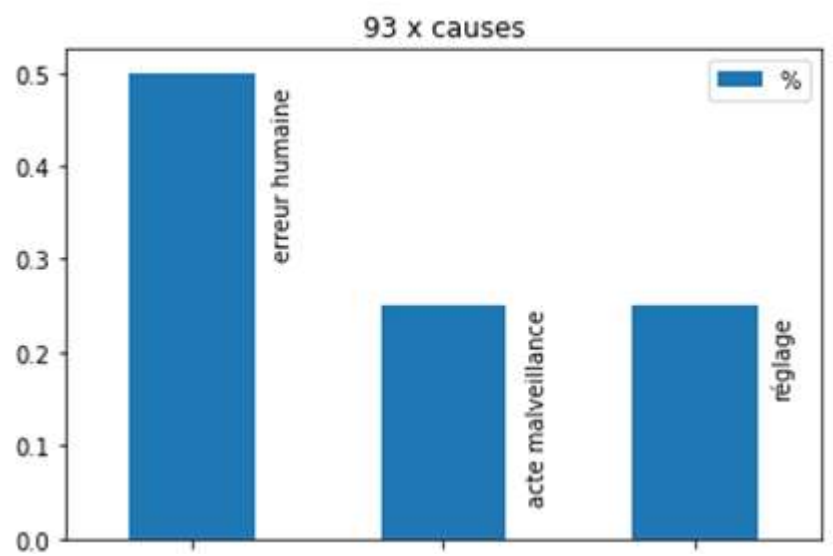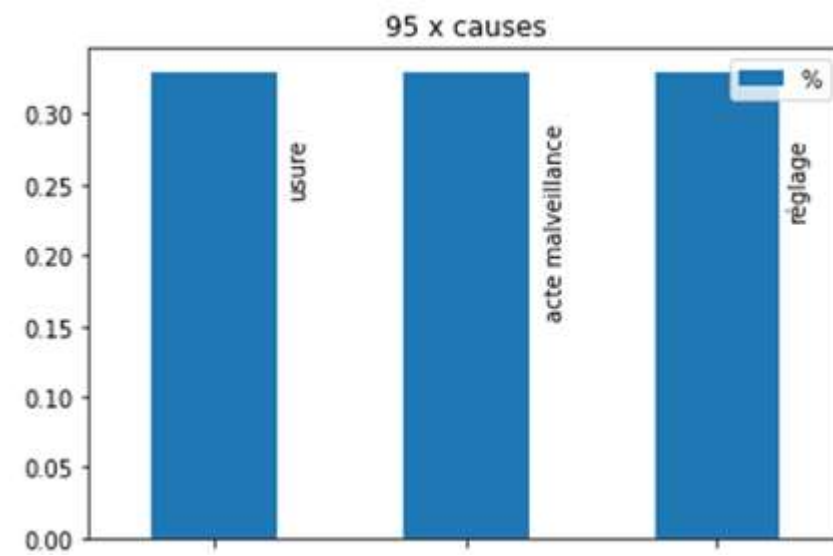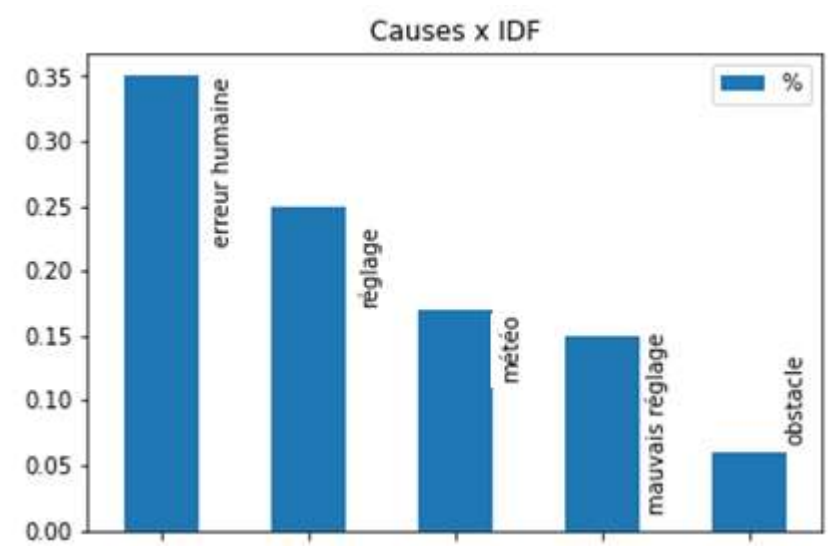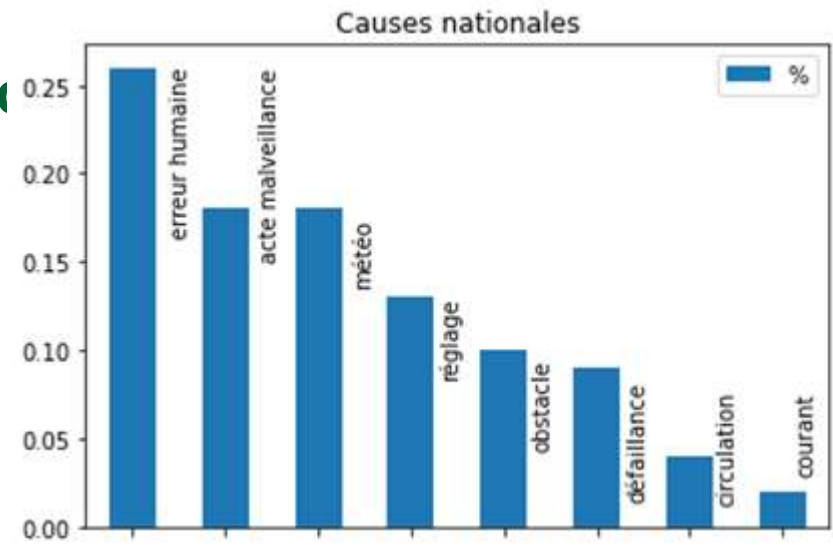
Example of efficiency for parsing