

Action MADONA, premiers résultats

Patrick Marcel, LIFAT, Université de Tours
Symposium Madics, 07/07/2021

MADONA

Maîtriser l'Analyse interactive de DONnées pour la NARRation journalistique



OÙ ?

Datajournalisme,
Data sciences,
Storytelling,
Communication
Journalistes,
Data scientistes
Associations citoyennes

POUR QUI ?



Journalistes
Services de communication
Start-ups de service Open data
Chargés de com'
Etudiants en infocom
Associations citoyennes

QUOI ?



Articles de la PQR et données sources
Données locales *Open data*
Données hétérogènes (structurées ou non)
Données *How-to* des journalistes

COMMENT ?



Enquêtes quali. sur les pratiques et usages
Réunions pluri-disciplinaires de validation des modèles
Hackathons datascientistes /datajournalistes

AVEC QUI ?

Laboratoire d'informatique fondamentale et Appliquée de Tours
Centre d'études sur les médias, les technologies et l'internationalisation
Centre de recherche sur les médiations
Rue89 Strasbourg

POUR QUOI ?



Livre blanc des pratiques
Prototype d'aide à l'analyse et la rédaction
Meilleure compréhension des processus de construction de narration



2019-2022

GARANTI SANS FAKE NEWS*!

<https://www.madics.fr/actions/madona/>

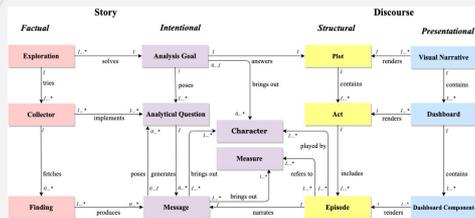
POUR QUOI ?



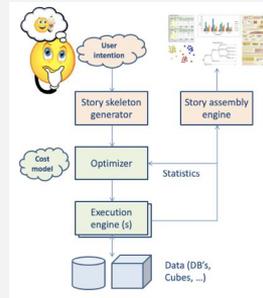
Meilleure compréhension des
processus de construction de
narration

Accompagner la mise en narration des données

- Objectif : Identifier les méthodologies de construction de narration à partir de données
 - Analyses des pratiques socio-professionnelles et analyses des productions
 - Automatisation de l'exploration de données et aide à la construction de narrations



Conceptualisation



Construire un vocabulaire

```
TAP Story for
This data analysis was automatically generated using the TAP algorithm.

In [1]: import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
import os
import sys
import time
import logging

# The data is already loaded. To reload it, use:
# reload_data()

# Parameters
year = 2012
country = 'USA'

# Data
data = pd.read_csv('data/usa_2012.csv')

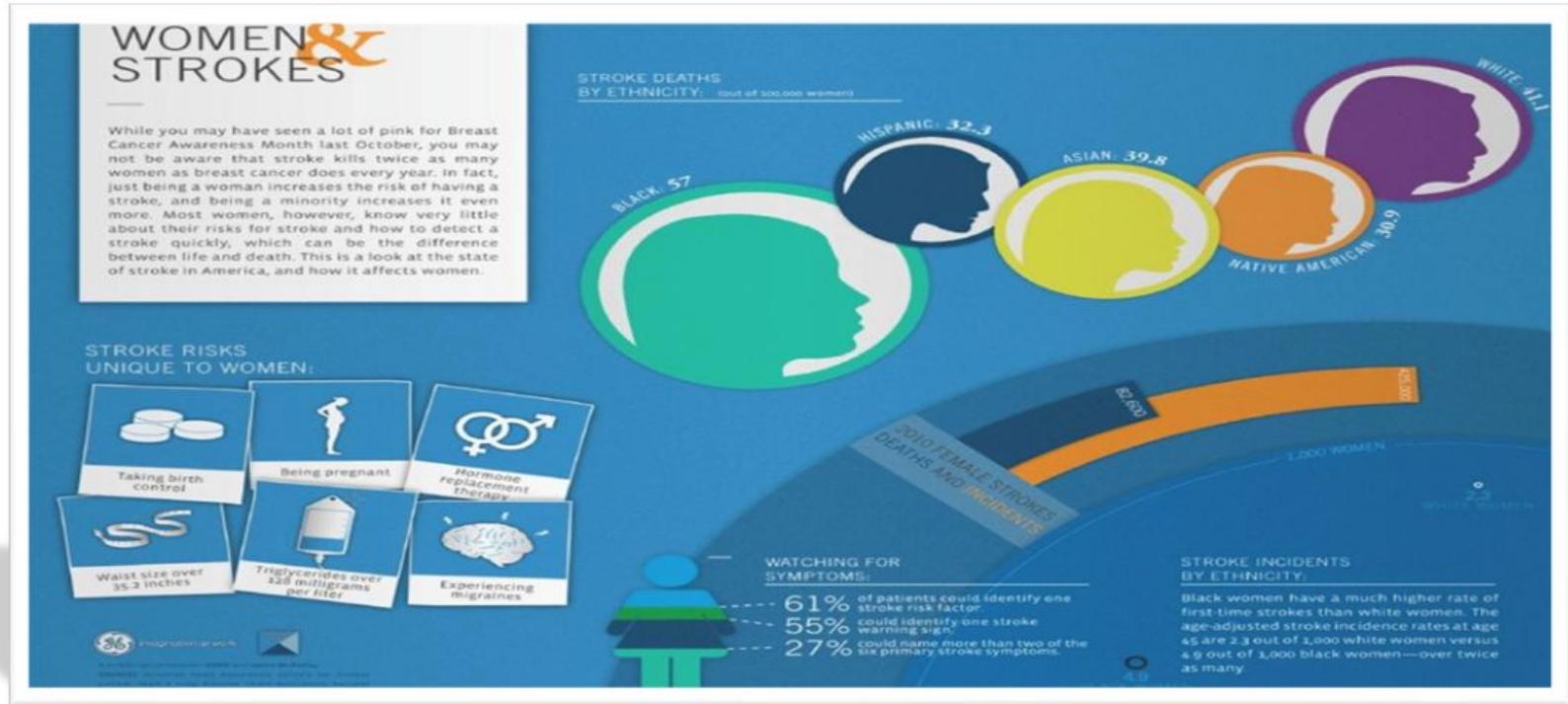
# Analysis
analysis = data.groupby('state').sum()

# Visualization
plt.figure(figsize=(10, 5))
sns.heatmap(analysis, cmap='YlOrRd', cbar=True)
plt.title('USA 2012')
plt.show()

# Output
print('USA 2012')
print(analysis)
```

Génération de notebooks

Exemple de narration de données



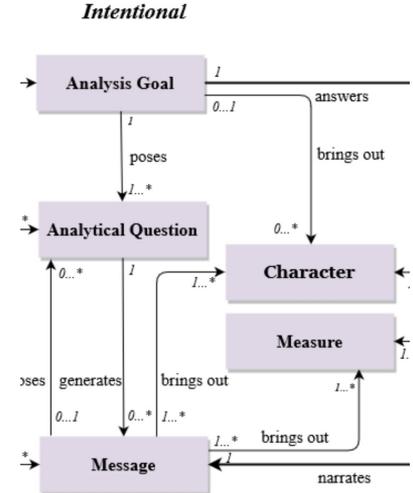
<https://www.good.is/infographics/facts-about-women-and-strokes>

Vocabulaire d'intentions

Simplifier l'accès et la manipulation de données

5 primitives

- **Description** (“**que** nous disent ces données ?”)
- **Comparaison** (“**comment** sont ces données par rapport à d'autres données ?”)
- **Explication** (“**pourquoi** ces données sont comme elles sont ?”)
- **Prédiction** (“comment seraient ces données **si...**?”)
- **Suggestion** (“que devrais-je savoir **d'autre** ?”)



Comparaison

2 prototypes de génération de notebooks de comparaisons :

1. Suite de comparatifs 1 contre 1, mise en évidence de faits statistiquement significatifs, prescriptif
2. Recommendation de comparaisons, mode automatique ou semi-automatique, nécessite un prétraitement en amont

```
TAP Story for  
This data analysis was automatically generated using the TAP algorithm.  
  
In [3]: import sqlalchemy  
sqlalchemy.create_engine("postgresql://localhost:5432/covid?user=marcel&password=159marcel")  
!load_ext sql  
!sql postgresql://localhost:5432/covid?user=marcel&password=159marcel  
!config sqlMagic.displaycon=False  
  
The sql extension is already loaded. To reload it, use:  
%reload_ext sql  
  
Q1  
Interestingness = 1.0  
Comparing year '2020' vs '2021' on Sum daily_vaccinations  
Correlation  
  
In [7]: !sql  
select t1.country,  
       t1.measure1 as "Sum(daily_vaccinations) for year = 2020", t2.measure2 as "Sum(daily_vaccinations) for year = 2021"  
from  
  (select year, country, sum(daily_vaccinations) as measure1  
   from covid_vac  
   where year = '2020'  
   group by year, country) t1,  
  (select year, country, sum(daily_vaccinations) as measure2  
   from covid_vac  
   where year = '2021'  
   group by year, country) t2  
where t1.country = t2.country;  
  
14 rows affected.  
  
Out [7]:  
country  Sum(daily_vaccinations) for year = 2020  Sum(daily_vaccinations) for year = 2021  
Czechia  10874  1028373  
Italy    13071  6397450  
Qatar    12158  324601  
Chile    17731  6281481  
Oman     4909  98494  
Belgium  404    110558
```



Intention 0
with sbora by nation
assess quantity

Intention 1
with benchmark & comparison suggestion
with sbora by nation
assess quantity
against REGION
using ratio(QUANTITY,
benchmark.QUANTITY)

Intention 2
with benchmark & comparison suggestion
with sbora by nation
assess quantity
against REGION scaled POPULATION
using ratio(QUANTITY,
benchmark.QUANTITY)

Intention 3
with benchmark & comparison suggestion
with sbora by nation
assess quantity
against REGION
using difference(QUANTITY,
benchmark.QUANTITY)

Intention 1 Labeling color: Neutral
Intention 2 Labeling color: Neutral
Intention 3 Labeling color: Neutral

Publications

1. Chagnoux M. (2020) "La datavisualisation, double point d'entrée du datajournalisme dans la PQR", Entre data journalisme et pratique infographique, Interfaces Numériques, Volume IX, n° 3/2020.
2. Faten El Outa, Matteo Francia, Patrick Marcel, Veronika Peralta and Panos Vassiliadis: [A conceptual model of data narrative for exploratory data analysis](#). ER 2020
3. Faten El Outa, Patrick Marcel, Veronika Peralta, Matteo Francia and Panos Vassiliadis: [Supporting the generation of data narratives](#). ER 2020 Demo. <https://github.com/OLAP3/pocdatastorytelling>
4. Alexandre Chanson, Ben Crulis, Nicolas Labroche, Patrick Marcel, Verónica Peralta, Stefano Rizzi, Panos Vassiliadis: [The Traveling Analyst Problem: Definition and Preliminary Study](#). DOLAP 2020
5. Antoine Chédin, Matteo Francia, Patrick Marcel, Verónica Peralta, Stefano Rizzi: The Tell-Tale Cube. [ADBIS 2020](#)
6. Verónica Peralta: From source data to data narratives: accompanying users in the way to interactive data analysis. Keynote, ADBIS/TPDL/EDA Joint Conferences, virtual conference, August 2020. https://youtu.be/A_tOQcyKc-w
7. Matteo Francia, Matteo Golfarelli, Patrick Marcel, Stefano Rizzi, and Panos Vassiliadis. Assess queries for interactive analysis of data cubes. [EDBT 2021](#)

Madona



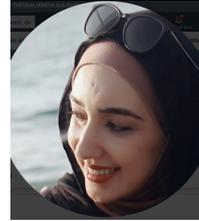
Marie



Patrick



Raphael



Faten



Alex



Lucile



Nicolas



Veronika



Julien



Matteo
(Bologna)



Panos
(Ioannina)



Stefano
(Bologna)

ANR Mobi'Kids (CNRS, univ. Rennes, Grenoble, Tours)

□ Objectifs

- Étude de l'autonomie des enfants et des cultures éducatives
- Développement d'un protocole méthodologique mixte
- Zone d'étude : **Rennes et Orgères**

□ Données

- Recueil de données des activités et des mobilités + données socio-psychologiques (entretien)
 - Tracker GPS + enquêtes sur tablette (activités, mode de déplacement, type d'accompagnement des enfants...)
 - 4 271 séquences d'activités journalières d'enfants de CM1, CM2 et 6^{ème} ou de l'un de leurs parents, recueil de 5 à 15 jours par an tous les 6 mois à 1-2 ans pour chaque individu
 - Données de parcours commentés
 - Données sensibles (CNIL)
- Reprise des EMD (enquête ménage-déplacement) de Rennes

□ Processus de fouilles de données

- Analyses des séquences d'activités journalières
- Définition de clusters de séquences similaires
- Explicabilité des clusters (éléments saillants de chaque clusters)
- ...

ANR – MOBI'KIDS, ANR-16-CE22-0009
(Coord. Sandrine Depeau)

ANR Mobi'Kids

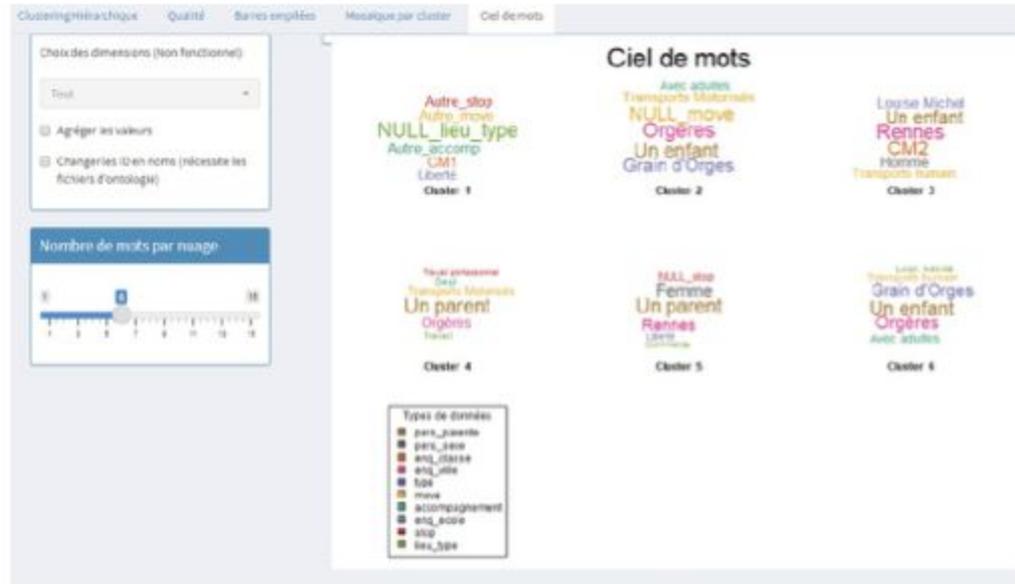
□ Développement d'un outil d'analyse et de visualisation : SIMBA

□ Exemple de connaissances extraites

- données Mobi'kids : Les comportements des enfants à Rennes et Orgères sont différents (mode de transport, accompagnement) :
- données EMD : Il y aurait un gain en autonomie lors du passage en 6^{ème} plus sensible pour les garçons que pour les filles (EMD Rennes)

□ Contacts

- thomas.devogele@univ-tours.fr
- sandrine.depeau@univ-rennes2.fr





**EGC 2022
BLOIS
24-28 janvier**
[**egc2022.univ-tours.fr**](http://egc2022.univ-tours.fr)



Sihem Amer-Yahia
Présidente du comité de
programme

Dates de soumission
Résumé : **8 octobre**
Article : **15 octobre**



<https://twitter.com/egc2022>