Université de **La Rochelle**
Laboratoire **Informatique Image Interaction**

Le projet NewsEye

Un enquêteur numérique pour la presse ancienne

Journée d'étude « A la recherche des femmes dans la presse ancienne numérisée (1850-1950) »

Antoine Doucet - 15 mars 2021

NEWS E E

NewsEye : Contexte du projet

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : contexte

Moyens et partenaires

Modalité

- Financement: 2 998 565€
- Parmi 3 projets sur 139 candidats (appel H2020 CULT-COOP-09-2017)
- Durée: 3 ans + 9 mois (01/05/2018 - 31/01/2022)
- Site: https://www.newseye.eu
- Suivez nous sur twitter: @NewsEyeEU
- La plateforme NewsEye: https://platform.newseye.eu/

4 pays, 5 langues principales, 3 types de partenaire (bibliothèque, SHS, informatique)

Université de la Rochelle:	Info
Bibliothèque nationale d'Autriche	Bibli
Université d'Helsinki	Bibli, SHS, Info
Université d'Innsbruck	SHS, Info
Bibliothèque nationale de France	Bibli
Université de Rostock	Info
Université Paul Valéry Montpellier 3	SHS
Université de Vienne	SHS

NEWS EYE

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : A Digital Investigator for Historical Newspapers

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Introduction

Motivation

- Presse ancienne: dizaines de millions de pages numérisées en Europe, nombreuses langues, nombreuses sources
- Une mine d'or pour plusieurs disciplines SHS et le grand public
  - 2500 utilisateurs/jour sur ANNO; chercheurs et grand public (enseignants/élèves, généalogistes, activités de loisir)
- Humanités numériques
  - Soutenir l'analyse qualitative par les données quantitatives,
  - Voire guider le qualitatif par le quantitatif

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Introduction

Constat

> Accès aux collections très difficile

- Contenus de mauvaise qualité
- Outils de recherche non adaptés à la presse ancienne

  De nombreux documents sont « manqués » (requêtes non exhaustives, problèmes d'OCR et de structuration)…
- Pas d'outils spécifiques pour l'analyse de la presse ancienne
- Une quantité d'information inadaptée à une analyse par les experts

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Défis

- Reconnaissance des articles et du contenu
  - > S’appuyer sur les articles plutôt que sur les pages

HUIT PAGES — Le Numéro quotidien (Paris et Départements) — CINQ CENTIMES

LE JOURNAL

RÉDACTION ET ADMINISTRATION : 100 RUE RICHELIEU, PARIS

FERNAND XAU, Fondateur

Prix des Abonnements

Annonces, Réclames et Faits Divers

Clinique des Délits

Les Révélations du “Times”

Les Chambres Anglaises écoutent des Déclarations fort édifiantes

ON NE PUBLIE PAS LA LETTRE DU KAISER

ÉCHOS

UN COMBAT AU MAROC

D’Amade met en déroute les Medakra et Hafidiens

L’ENNEMI SUBIT D’ÉNORMES PERTES

Le Rapport Officiel

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Défis

- Reconnaissance des articles et du contenu
  - > S’appuyer sur les articles plutôt que sur les pages

HUIT PAGES — Le Numéro quotidien (Paris et Départements) — CINQ CENTIMES

LE JOURNAL

RÉDACTION ET ADMINISTRATION : 100 RUE RICHELIEU, PARIS

FERNAND XAU, Fondateur

Clinique des Délits

Les Révélations du “Times”

Les Chambres Anglaises écoutent des Déclarations fort éloquentes

ON NE PUBLIE PAS LA LETTRE DU KAISER

UN COMBAT AU MAROC

D’Amade met en déroute les Medakra et Hafidiens

L’ENNEMI SUBIT D’ÉNORMES PERTES

Le Rapport Officiel

ÉCHOS

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Défis

- Reconnaissance des articles et du contenu
  - > S’appuyer sur les articles plutôt que sur les pages
  - > Extraire le contenu textuel

HUIT PAGES — Le Numéro quotidien (Paris et Départements) — CINQ CENTIMES

LE JOURNAL

RÉDACTION ET ADMINISTRATION : 100 RUE RICHELIEU, PARIS

FERNAND XAU, Fondateur

Clinique des Délits

Les Révélations du “Times”

Les Chambres Anglaises écoutent des Déclarations fort édifiantes

ON NE PUBLIE PAS LA LETTRE DU KAISER

UN COMBAT AU MAROC

D’Amade met en déroute les Medakra et Hafidiens

L’ENNEMI SUBIT D’ÉNORMES PERTES

Le Rapport Officiel

ÉCHOS

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Défis

- Reconnaissance des articles et du contenu
  - > S’appuyer sur les articles plutôt que sur les pages
  - > Extraire le contenu textuel

HUIT PAGES — Le Numéro quotidien (Paris et Départements) — CINQ CENTIMES

LE JOURNAL

RÉDACTION ET ADMINISTRATION : 100 RUE RICHELIEU, PARIS

FERNAND XAU, Fondateur

Clinique des Délits

Les Révélations du “Times”

Les Chambres Anglaises écoutent des Déclarations fort décevantes

ON NE PUBLIE PAS LA LETTRE DU KAISER

UN COMBAT AU MAROC

D’Amade met en déroute les Medakra et Hafidiens

L’ENNEMI SUBIT D’ÉNORMES PERTES

Le Rapport Officiel

ÉCHOS

Reconnaissance Optique de Caractères

Les Révélations
du "Times"
Les Chambres Anglaises écoutent
des Déclarations fort démâtes
01KE PUBLIE PAS LA LETTRE DU KAISER
1,'opînion, en Angleterre, et aussi un peu ailleurs, attendait avec une vive curiosité les explications qui devaient être apportées hier aux deux Chambres du Parlement britannique, au sujet de la correspondance éehaagée entre Guillaume II et lard Tweedmou ttju Sera-t-elle satisfaite œes très brèves déclarations faites par le premier tord de l'Amirauté et par M. Asqniîb. ? Il est difficile de le croire.

The NewsEye project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Défis

- Reconnaître les entités nommées
  - > Personnes, lieux, organisations
  - > Pour une analyse plus précise, pour construire des corpus
  - > Ouverture à une analyse quantitative

HUIT PAGES — Le Numéro quotidien (Paris et Départements) — CINQ CENTIMES

LE JOURNAL

RÉDACTION ET ADMINISTRATION : 100 RUE RICHELIEU, PARIS

FERNAND XAU, Fondateur

Clinique des Délits

Les Révélations du "Times"

Les Chambres Anglaises écoutent des Déclarations fort décevantes

ON NE PUBLIE PAS LA LETTRE DU KAISER

UN COMBAT AU MAROC

D'Amade met en déroute les Medakra et Hafidiens

L'ENNEMI SUBIT D'ÉNORMES PERTES

Le Rapport Officiel

ÉCHOS

Reconnaissance Optique de Caractères

Les Révélations
du "Times"
Les Chambres Anglaises écoutent
des Déclarations fort démâtes
01KE PUBLIE PAS LA LETTRE DU KAISER
1,'opînion, en Angleterre, et aussi un peu ailleurs, attendait avec une vive curiosité les explications qui devaient être apportées hier aux deux Chambres du Parlement britannique, au sujet de la correspondance éehaagée entre Guillaume II et lard Tweed-mou ttju Sera-t-elle satisfaite œes très brèves déclarations faites par le premier tord de l'Amirauté et par M. Asqniîb. ? Il est difficile de le croire.

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Défis

- Identifier les entités nommées
  - > Quel Washington ? Quel président de la république ?
  - > Connexion à des bases de connaissances externes
  - > Ouverture à une analyse cross-lingue

HUIT PAGES — Le Numéro quotidien (Paris et Départements) — CINQ CENTIMES

LE JOURNAL

RÉDACTION ET ADMINISTRATION : 100 RUE RICHELIEU, PARIS

FERNAND XAU, Fondateur

Clinique des Délits

Les Révélations du "Times"

Les Chambres Anglaises écoutent des Déclarations fort décevantes

ON NE PUBLIE PAS LA LETTRE DU KAISER

UN COMBAT AU MAROC

D'Amade met en déroute les Medakra et Hafidiens

L'ENNEMI SUBIT D'ÉNORMES PERTES

Le Rapport Officiel

ÉCHOS

Reconnaissance Optique de Caractères

Les Révélations
du "Times"
Les Chambres Anglaises écoutent
des Déclarations fort démâtes
01KE PUBLIE PAS LA LETTRE DU KAISER
1,'opînion, en Angleterre, et aussi un peu ailleurs, attendait avec une vive curiosité les explications qui devaient être apportées hier aux deux Chambres du Parlement britannique, au sujet de la correspondance éehaagée entre Guillaume II et lard Tweed-mou ttju Sera-t-elle satisfaite œes très brèves déclarations faites par le premier tord de l'Amirauté et par M. Asqniîb. ? Il est difficile de le croire.

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Objectifs

Ambitions

> Faciliter l'accès aux collections de presse ancienne
  - Réduire le bruit (OCR, structuration)
  - Enrichir les contenus sémantiquement
  - Inventer des méthodes robustes au bruit
  - Inventer des approches indépendantes des langues

> Permettre une analyse quantitative…
  - qui soit justifiable et interprétable de façon qualitative

> Impacter tous les utilisateurs potentiels de la presse historique
  - Grand public, enseignants, journalistes, bibliothécaires, étudiants, …
  - Chercheurs en SHS (histoire, littérature, linguistique, géographie, arts du spectacle…)

> **Toute langue, tout corpus, toute discipline**
  - immédiatement applicable à tout autre corpus (**approche durable**)

> Contribuer à la recherche en SHS **et en informatique**

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Ressources

Collections

- Période : 1850-1950
- Jeux de données : 15M pages (bibliothèques nationales de 3 états)
- Entièrement analysés : 1.5M pages
- Objectif clé : permettre les extensions
  - Intégrer de nouvelles sources
  - Intégrer de nouveaux langages
  - Intégrer de nouveaux outils

Au-delà des cas d'utilisation

- Cas d'utilisation pour estimer le potentiel de NewsEye pour la recherche en sciences humaines
  - Littérature / Français
  - Histoire / Finnois-Suédois
  - Histoire / Allemand
  - Journalisme / Allemand

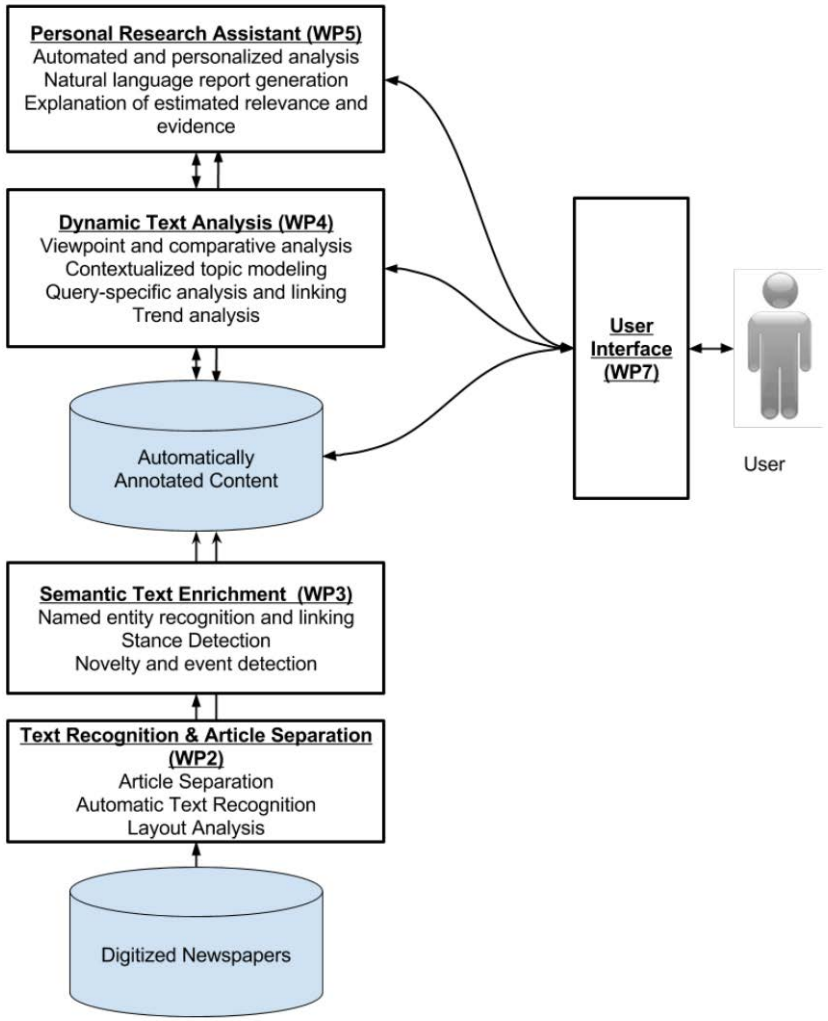
  [possibilité de franchir les frontières linguistiques : tous]

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : présentation

Organisation

- Au delà des livrables:
  - > Production itérative et incrémentale
  - > Accès en continu
- Pour les partenaires SHS
  - > Accès et retour de pertinence en continu
  - > Pas de phase d'attente

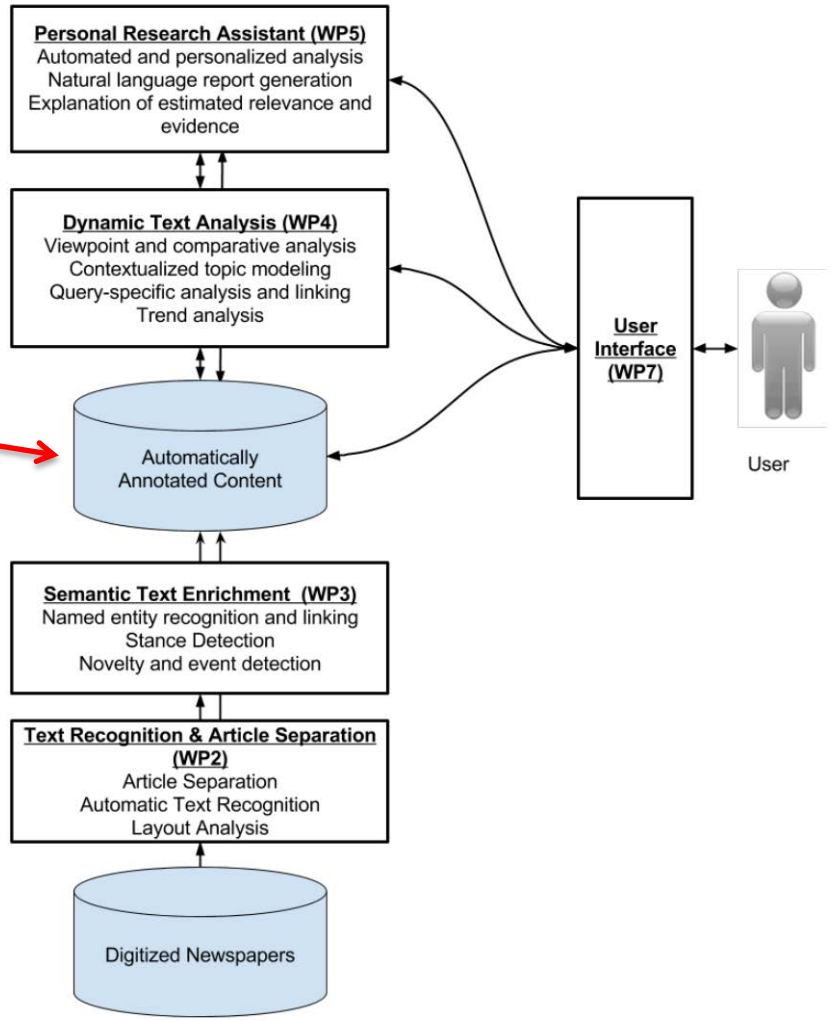


The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : Enrichissement sémantique

Résultat

- Une base de connaissances entièrement indépendante des langues
- Des outils pour y accéder/rechercher
- D'autres outils pour découvrir, enquêter, rapporter et expliquer

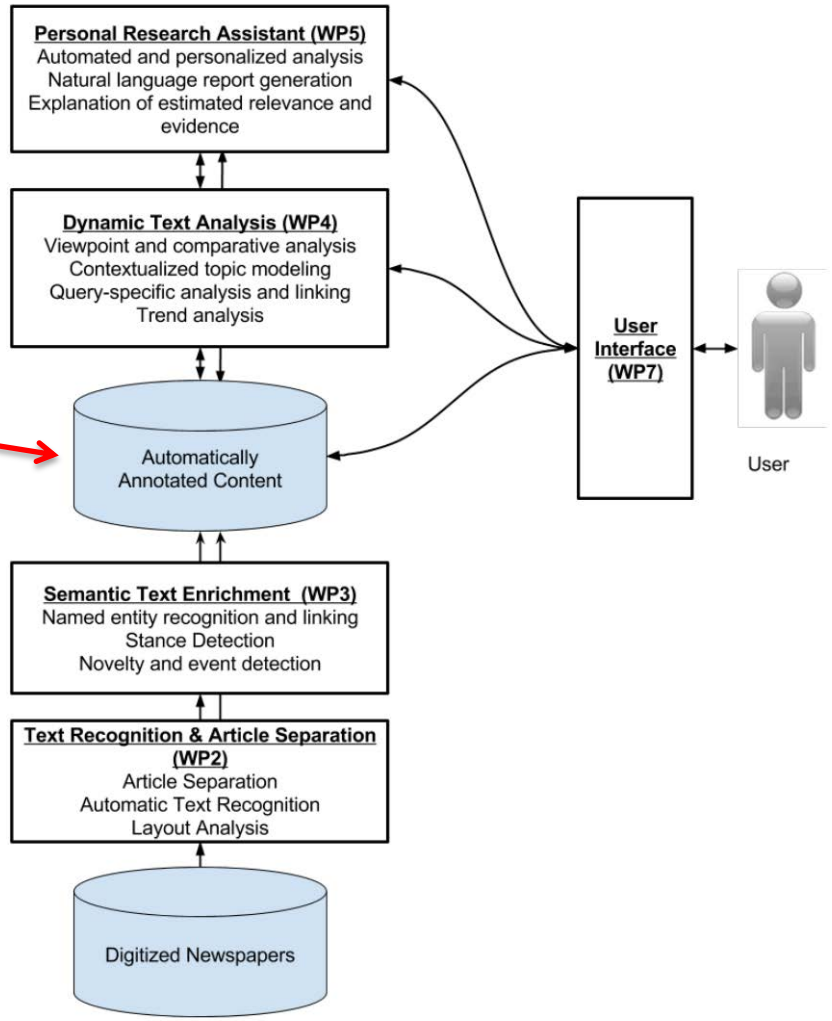


The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : Enrichissement sémantique

Résultat

- Une base de connaissances entièrement indépendante des langues
- Des outils pour y accéder/rechercher
- D'autres outils pour découvrir, enquêter, rapporter et expliquer



- Reconnaissance d'entités nommées
  - > Compétition internationale HIPE@CLEF 2020
  - > Noms de personnes, de lieux, d'organisations
  - > Meilleur système dans 50 évaluations sur 52
    - en français, anglais, allemand

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

A la recherche des femmes dans la presse anciennes numérisée (1850-1950)

Où sont les femmes ?

> Approches qualitatives, comme l'exploitation des occurrences de prénoms et pronoms féminins et masculins
  - Fréquence relative comparée
  - Contexte d'occurrence (proéminence, analyse d'opinion)
  - Possibilité de franchir les frontières linguistiques
> Méthodes plus qualitatives utilisant par exemple le *topic modeling* où la détection d'évènements
  - S'appuyant sur des évènements significatifs (droit de vote, journée internationale des femmes)
  - Perception de ces moments clés dans la presse
  - Analyse comparative dans différents pays et médias
> Méthodes entièrement qualitatives et/ou manuelles
  - Construction et export de collections
  - Utilisation d'outils externes et analyse manuelle

Les femmes et les pantalons

> Sujet de recherche de l'université Paul-Valéry-Montpellier

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye : la plateforme

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

NewsEye

Plateforme – à vous de jouer

A vous de jouer !

- Site : https://www.newseye.eu
- Suivez nous sur twitter : @NewsEyeEU
- **La plateforme NewsEye : https://platform.newseye.eu/**

Tout est disponible :

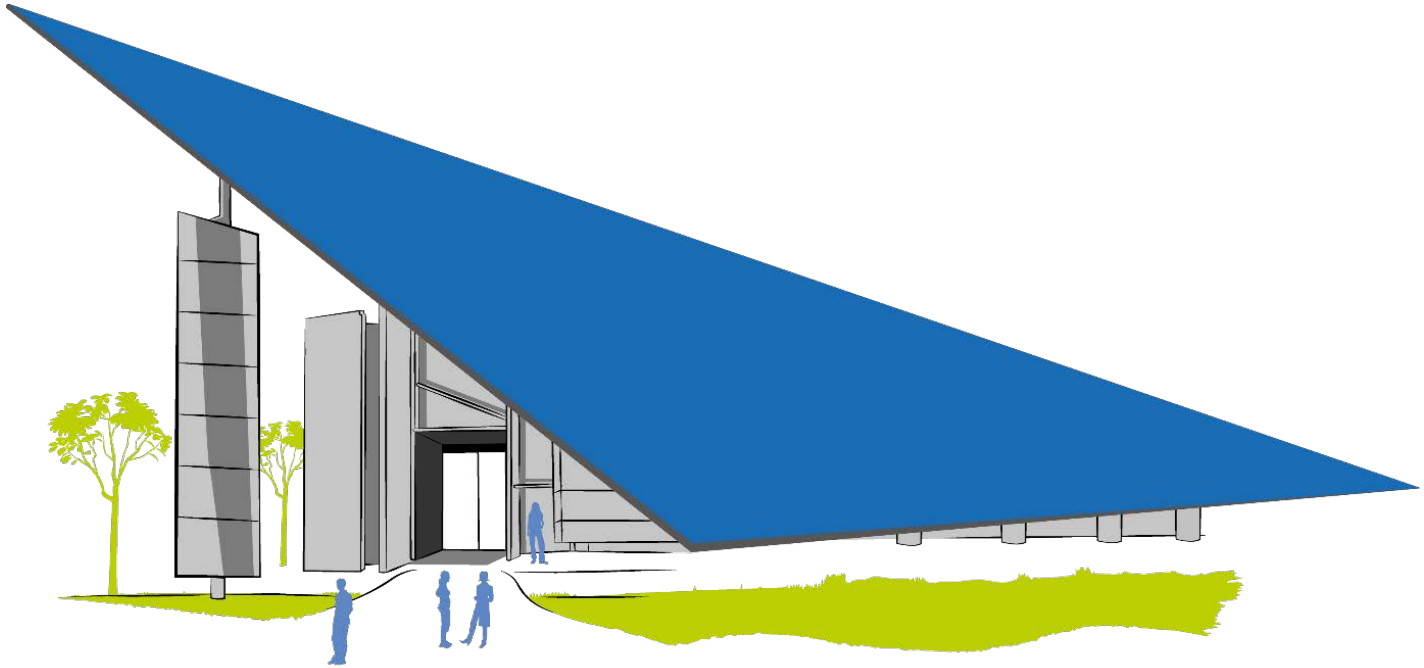
- Articles scientifiques et jeux de données : https://zenodo.org/communities/newseye/
- Code source : https://github.com/NewsEye/
- Matériel pédagogique : https://www.newseye.eu/fr/ressources/materiel-pedagogique/
- Et bien plus encore !

Collaborations bienvenues

- La plateforme atteint sa phase d'exploitation

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.

Université de **La Rochelle**



L'Université nouvelle génération

23 avenue Albert Einstein - BP 33060 - 17031 La Rochelle

www.univ-larochelle.fr