

Fouille de texte et enrichissement sémantique de la presse ancienne européenne

Antoine Doucet – 6 Décembre 2018

NewsEye

Contexte

➤ Projet H2020

- > Durée : 3 ans (mai 2018 – avril 2021)
- > Web : <https://www.newseye.eu>
- > Twitter : @NewsEyeEU

➤ 4 pays, 5 langues principales, 3 types de partenaires (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 |

Fouille et enrichissement de journaux anciens

Introduction

➤ Motivation

- > Presse historique: 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



Fouille et enrichissement de journaux anciens

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



NewsEye Objectifs

➤ Ambition

- > Faciliter l'accès aux collections de presse ancienne
 - Réduire le bruit (OCR, structuration)
 - Enrichissement sémantique
 - Créer une approche indépendante des langues, robuste au bruit
- > Permettre une analyse quantitative **compréhensible**, qui soit **justifiable** et **explicable** par des pointeurs qualitatifs
- > Impacter tous les utilisateurs de la presse ancienne
 - Grand public, enseignants, élèves, journalistes, bibliothécaires, étudiants...
 - Chercheurs en histoire, littérature, linguistique, arts du spectacle, ...
- > **Toute langue, tout corpus, toute discipline**
 - Possibilité d'intégrer tout autre jeu de données (**approche durable**)
- > Contribuer à la recherche **en SHS et en informatique**

NewsEye

Données initiales

➤ Corpus

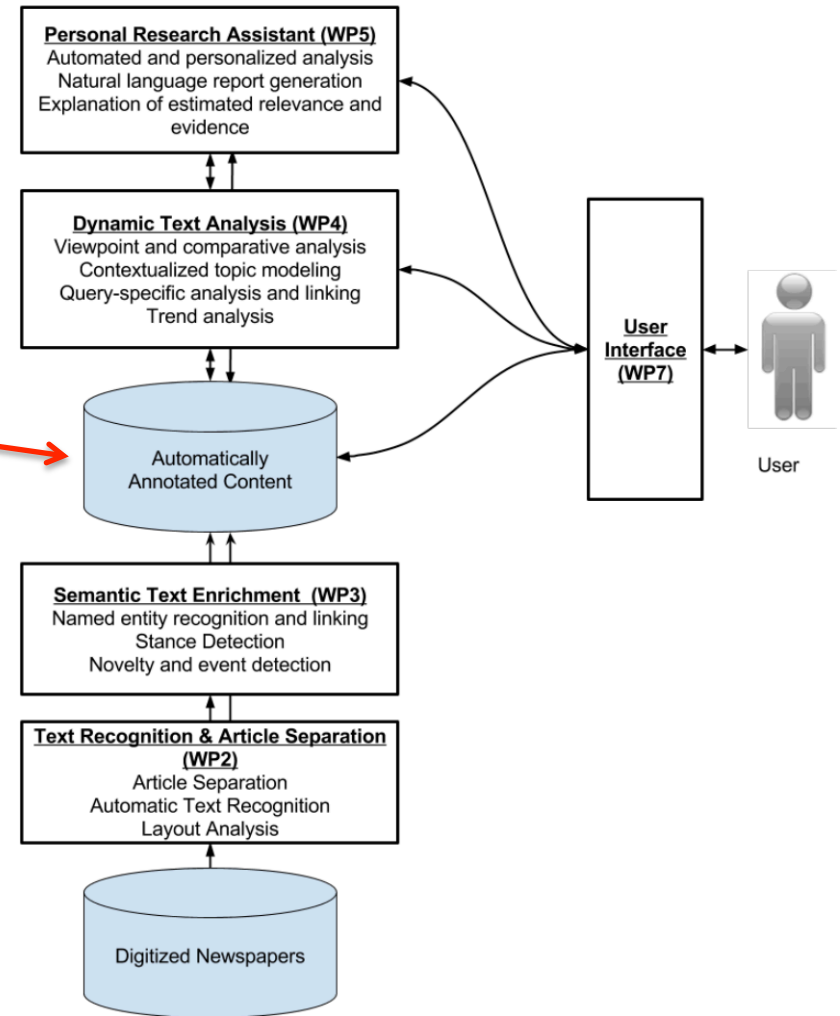
- > Période de focus du projet : 1850-1950
- > Jeu de données : 15 M pages (des 3 bibliothèques nationales partenaires)
- > Tous les processus sont ouverts aux extensions
 - Plus de données/sources
 - Plus de cas d'utilisation
 - Extensions logicielles

➤ Au-delà des cas d'utilisation

- > Les sujets de recherche SHS du projet sont des cas d'utilisation en vue de démontrer le potentiel du projet pour la recherche en sciences humaines
 - Littérature / français
 - Histoire / finnois-suédois
 - Histoire / allemand
 - Journalisme / allemand
- [capacité à franchir les frontières linguistiques: tous]

Enrichissement sémantique Processus

- Une base de connaissances indépendante des langues
- Des outils pour y accéder
- Des outils pour y découvrir, investiguer, rapporter et expliquer



Enjeux science ouverte

ORDP

➤ *Open Research Data Pilot* (optionnel)

- > Tout est à spécifier d'emblée (et m à j continue)
- > *Data Management Plan*
- > Requière d'être FAIR:
 - *Findable*
 - *Accessible*
 - *Interoperable*
 - *Reusable*

➤ Oui ! Mais...

Enjeux science ouverte

ORDP

➤ *Open Research Data Pilot* (optionnel)

- > Tout est à spécifier d'emblée (et m à j continue)
- > *Data Management Plan*
- > Requière d'être FAIR:
 - *Findable*
 - *Accessible*
 - *Interoperable*
 - *Reusable*

➤ Oui ! Mais...

➤ *Findable+Interoperable*: OK

➤ *Accessible+Reusable*: pas si simple

- > Publications: aucun problème
- > Données et outils: distinction entre les entrées et les sorties
 - Pour les données comme pour les logiciels

Enjeux science ouverte

FAIR

➤ Logiciel

- > Sujet aux licences antérieures
- > Nos licences ne peuvent pas être plus permissives que celles des outils que nous utilisons...

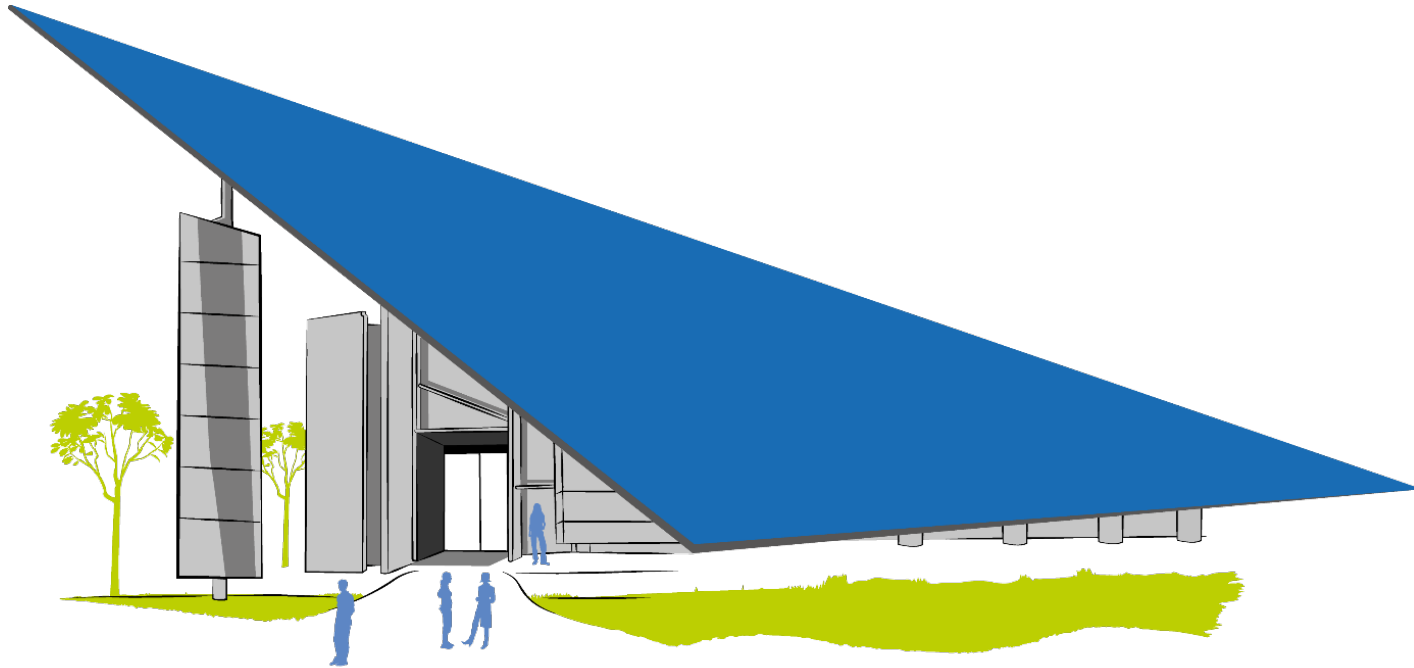
➤ Données/Corpus

- > OK pour ce que le projet produit
- > Mais le matériau initial a son propre copyright
 - Autriche; uniquement sur son propre portail
 - Finlande: ça dépend de l'année de publication
 - France: ça dépend de l'année et du cadre de numérisation
- > Approche globale difficile
- > Une zone grise: Données personnelles et ayant-droits...

➤ Enrichissement accessible d'un texte qui ne l'est pas !?

- > Solution: redistribution des productions aux bibliothèques

University of La Rochelle



L'Université nouvelle génération

23 avenue Albert Einstein - BP 33060 - 17031 La Rochelle

www.univ-larochelle.fr