



Numérisation des cahiers de doléance

Long time no see ... J'écris surtout ailleurs depuis quelques années, mais là, petit projet perso qui se fait jour.

Suite à des discussions sur les réseaux sociaux après les divers coupures de presses parues à propos des cahiers de doléances du « grand débat » d'Emmanuel Macron qui semble s'enliser dans les archives départementales, ça me semblerai pas mal de tenter un docjacking de tout ça pour le numériser et le publier nous même, puisque nos amis de la stratosphère trouvent que ça fait trop de téraoctets.

Appel à bonnes volontés, donc.

Il faut, vu de mon petit bout de lorgnette :

- Constituer des groupes départementaux qui iront scanner l'ensemble des documents (coucou les GJ ! Vous voulez aider ?)
- Rentrer tout ça dans une plateforme de transcription qu'il faudra probablement créer (on trouvera sans trop de soucis ou faire héberger tout ça)
- Répartir les rôles entre des personnes effectuant la transcription, d'autres validant ces transcriptions et un dernier rôle consistant à taguer les transcriptions pour pouvoir les ranger plus facilement après

Il serait bien entendu plus simple de récupérer ce que la BnF a déjà passé au scanner (soit, si on en croit les journaliste, la totalité) voir même ce qui a déjà été transcrit .. Mais sauf à faire grand bruit, ça semble peu probable, et une partie de l'exercice consiste surtout à ce que plein de gens lisent ce que plein de gens ont écrit. Quoi de mieux qu'une transcription pour ça ?

Tu cause à des élus locaux ou départementaux (voir t'es élu toi même) ? Tu sais développer ? Tu as déjà joué dans le monde de la transcription de documents ? Tu as des infrastructures d'hébergement avec des téraoctets dispos ? Tu sais lire et taper au clavier ? T'as un scanner et un ordinateur portable ? Tu trouve juste ça rigolo et tu veux y mettre les mains ? Viens aider :)

Étape en cours : aller dénicher un cahier voir comment il est fichu, ce qu'on y trouve, ce qui doit être enlevé (anonymisation), comment ça peut se scanner facilement, comment documenter la partie scanning et préparer le développement de la plateforme de transcription.

Problématique de l'anonymisation : une partie des écrits doivent immanquablement contenir des données personnelles (que ce soit l'adresse de la personne qui a rédigé, ou le contenu en lui même qui peut faire mention de personnes, d'entreprises, etc ...). Il faudra veiller à leur anonymisation.

Problématique de la pollution : lorsqu'on crowdfund une opération, il y a nécessairement des plaisantins qui viennent polluer le travail des autres. Ainsi, une transcription devra être validée



par au moins 2 ou 3 personnes, si possible géographiquement éloignées du transcripateur lui-même. Un versioning poussé des transcription sera par ailleurs nécessaire pour éviter tout attaque type defacing.