INTD - Titre 1 - Novembre 2013 - Projet veille

Livrable 2 : janvier 2014

Consignes pour la préparation du corpus destiné à un traitement DataMining (à réaliser avant le 5/12/2013)

Le corpus pourra être constitué par exemple d'articles de presse, des dépêches d'agence (ou news >>> google), des actus/news d'un réseau social (pas twitter), des outils de curation (scoop it de divers auteurs), de blogs, de pages web de sources institutionnelles, politiques, associatives etc.

Taille des documents à copier/coller : 2 pages maximum.

Volume de documents: 150-200

Les méta-données à indiquer pour chaque document, et afin de réaliser des traitements permettant de travailler à l'analyse de clusters sémantiques et sociaux (qui dit quoi), en étudier l'évolution dans le temps =

* Date de création du document (mois et/ou année selon périmètre temporel traité)

* Source : institution, organisme, titre du support ou du site etc.

* Auteurs

* url

* Pays (si requis par le sujet)

Fichier de traitement : en word ou Excel/CSV

Les champs doivent être séparés par une balise (si format word) ou sous forme de tableau (excel ou csv) avec

- une colonne pour chaque méta-donnée
- une colonne pour le champ texte

Les fichiers doivent être enregistrés au format txt brut (si modèle .doc) et au format txt ou CSV avec séparateur tabulation (si modèle xlsx).

Le corpus doit être prêt pour le 5 décembre.