ANR CORPUS GEOMEDIA
Sample Agenda

These data come from the GEOMEDIA database, created within the ANR CORPUS GEOMEDIA project. This database collects and stores data sent by media websites through RSS feed technology. Currently, 300 RSS feeds from 161 different media are collected permanently, in 8 languages and from 59 different countries. At the end of the project (June 2016), data should be available in free access. To initiate exchanges with the scientific world about treatment, enrichment and visual projection methods, we decided to provide an early sample of the database.

USE OF DATA IS ONLY ALLOWED IN A SCIENTIFIC CONTEXT.

DATA CANNOT BE USED FOR COMMERCIAL PURPOSES.

In this sample, we provide all the items (news) send by 36 RSS feeds, from the 28th of April, 2014, to the 26th of April, 2015. These RSS feeds are called « international » because they are categorized as follow on their website of origin: « international », « mundo », « internacional », « monde », « world news » or « world ».

I. List of RSS feeds provided in the sample :

TO ADD: annex of our paper in *L'Esapce Géographique*.

II. Presentation of each data files provided by feed

For each feed, we provide 3 data files and one statistic report:

- **rss.csv** = the whole raw data
- **rss_unique.csv** = data without duplicated items (cf. next part)
- **rss_unique_TAG_country_Ebola.csv** = data without duplicated items, and tagged by countries and regarding some Ebola-related keywords (cf. next part).

a. rss.csv

This file contains all the raw data collected, which means the entire items (news) collected from the 28th of April, 2014, to the 26th of April, 2014 for each feed.

The file is structured in 5 fields:

Name	Type	Definition
ID	Character	Unique key of the item.
feed	Character	Unique code of the feed.
time	Date	Day & hour of item collection. It is not the time of item publication but of its recovery. The gathering tool works permanently, and tries to recover items for each feeds on an hourly basis. Therefore the time difference between publication & collection time should not be really significant.
text1	Character	Title of RSS item. In RSS case, it should be the title of an article.
text2	Character	Description of RSS item. In RSS case, it should be a part of the (or the whole) article.

b. rss_unique.csv

Geomedia database automatically deletes duplicated items from the database: if a collected item is strictly identical to another already stored (same feed), the app does not store the last item issued. But, if only a small part of an item has been modified (orthographic & punctuation correction), the app will store the item a second time without detecting the duplication. It is the reason why, in this sample, we also provide « clean data », called « rss_unique.csv ».

In this file you will find the same data as in « rss.csv » without duplicated items. We have deleted all the items than have a strictly identical title (text1) OR identical description (text2) over a seven day period.

This file contains all the unique items collected from the 28th of April, 2014, to the 26th of April, 2015, for each feeds. The file structure is exactly the same as in the previous file.

If you merge « rss.csv » and « rss_unique.csv » by the ID, you can detect all deleted items.

c. rss_unique_TAG_country_Ebola.csv

In the context of the GEOMEDIA project, corpus enrichment is one of the main data treatment that has been done. We have geo-tagged items with the countries quoted **in the title** (text1). To do it, we built a word dictionary which allows to automatically detect countries. In this sample, items have also been tagged with key-words related to the Ebola virus (to illustrate thematic experiments that can been done from the data). **The two dictionaries used to tag the data are also provided (Dico_Country_Free.csv and Dico_Ebola_Free.csv).**

In this file, you will find all the items of « rss_unique.csv », geo-tagged with the dictionary. The file structure is exactly the same as for previous files although it contains two additional fields.

File is structured in 7 fields:

Name	Type	Definition
ID	Character	Unique key of item.
feed	Character	Unique code of feed.
time	Date	Day & hour of item collection. It is not the time of item publication but of its recovery. The gathering tool works permanently, and tries to recover items for each feeds on an hourly basis. Therefore the time difference between publication & collection time should not really significant.
text1	Character	Title of RSS item. In RSS case, it should be the title of an article.
text2	Character	Description of RSS item. In RSS case, it should be a part (or the entire) of the article.
TAG_country	Character	ISO3 code of countries detected in the item title (text1). Empty if none.
TAG_ebola	Character	Equal “Ebola” if a related key-word has been detected in the title (text1). Empty else.

If several countries have been detected, items are duplicated as many times as the quoted number of countries. Example:

ID	Feed	time	Text1	TAG_country
3117022	en_AUS_austra_int	2014-10-01 00:33:26	Hong Kong protesters vow to stay put	**HKG**
3117501	en_AUS_austra_int	2014-01-01 15:45:32	India puts US diplomats on notice	**IND**
3117501	en_AUS_austra_int	2014-01-01 15:45:32	India puts US diplomats on notice	**USA**
3117502	en_AUS_austra_int	2014-01-01 00:43:12	Rebel boss talks of peace but fights on	
3117503	en_AUS_austra_int	2014-10-01 01:33:27	Secret Service lashed for Obama breach	**USA**

d. statistic report (html)

With the data, we also provide some statistic reports (html format) for each feed.

These reports are being developed in the context of the project. **The development is still in progress, and is currently restricted to French.**

These reports present many indicators and several simple visual representations of the “raw data”, “unique data” and “tagged data”. **It enables to have a better idea of the available data in this sample, and give some example of feasible visual representation.**

Contact us: **geomedia@gis-cist.fr**