ANR Corpus Geomedia

Free sample of the RSS database

Data come from the Geomedia database, created as part of ANR corpus Geomedia. 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 8 RSS feeds, from 1st October to 31 December 2014. 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 :

- *Daily Newspaper:* **The Australian**
  Feed category: International
  Language: **English**
  Origin country: **Australia**
  Database ID: en_AUS_austra_int
  Web site URL: http://www.theaustralian.com.au
  Feed URL: http://timesofindia.feedsportal.com/c/33039/f/533917/index.rss

- *Daily Newspaper:* **El Mercurio**
  Feed category: International
  Language: **Spanish**
  Origin country: **Chile**
  Database ID: es_CHL_mercur_int
  Web site URL: http://www.emol.com
  Feed URL: http://timesofindia.feedsportal.com/c/33039/f/533917/index.rss

- *Daily Newspaper:* **South China Morning Post**
  Feed category: International
  Language: **English**
  Origin country: **China (Hong Kong)**
  Database ID: en_CHN_mopost_int
  Web site URL: http://www.scmp.com
  Feed URL: http://timesofindia.feedsportal.com/c/33039/f/533917/index.rss

- *Daily Newspaper:* **Le Monde**
  Feed category: International
  Language: **French**
  Origin country: **France**
  Database ID: fr_FRA_lmonde_int
  Web site URL: http://www.lemonde.fr
  Feed URL: http://rss.lemonde.fr/c/205/f/3052/index.rss

- *Daily Newspaper:* **The Daily Telegraph**
  Feed category: International
  Language: **English**
  Origin country: **United Kingdom**
  Database ID: en_GBR_dailyt_int
  Web site URL: http://www.telegraph.co.uk
  Feed URL: http://www.telegraph.co.uk/news/worldnews/rss

- *Daily Newspaper:* **The Times of India**
  Feed category: International
  Language: **English**
  Origin country: **India**
  Database ID: en_IND_tindia_int
  Web site URL: http://timesofindia.indiatimes.com
  Feed URL: http://timesofindia.feedsportal.com/c/33039/f/533917/index.rss

- *Daily Newspaper:* **El Universal**
  Feed category: International
  Language: **Spanish**
  Origin country: **Mexico**
  Database ID: es_MEX_univer_int
  Web site URL: http://www.eluniversal.com.mx
  Feed URL: http://www.eluniversal.com.mx/rss/mundo.xml

- *Daily Newspaper:* **The New York Times**
  Feed category: International
  Language: **English**
  Origin country: **United States of America**
  Database ID: en_USA_nytime_int
  Web site URL: http://www.nytimes.com
  Feed URL: http://www.nytimes.com/services/xml/rss/nyt/World.xml

Summary table of RSS feeds:

Daily NewsPaper	Feed category	Language	Country	Website	Total items collected 1 Oct to 31 Dec 2014
The australian	International	English	**Australia**	http://www.theaustralian.com.au	1 493
El mercurio	International	Spanish	**Chile**	http://www.emol.com	3 046
South China Morning Post	International	English	**China**	http://www.scmp.com	1 975
Le Monde	International	French	**France**	http://lemonde.fr	3 998
Daily telegraph	International	English	**United Kingdom**	http://www.telegraph.co.uk	5 609
The times of India	International	English	**India**	http://timesofindia.indiatimes.com	1 427
El Universal	International	Spanish	**Mexico**	http://www.eluniversal.com.mx	4 167
The New York Times	International	English	**United states**	http://www.nytimes.com	5 871

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_tagged.csv** = data without duplicated items, and tagged by countries (cf. next part).

a. rss.csv

This file contains all the raw data collected, which means the entire items (news) collected from 1st October to 31 December 2014 for each feed.

File is structured in 5 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 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 (orthograph & 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 1st October to 31 December 2014 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_tagged.csv

In the context of ANR Geomedia, corpus enrichment is one of the main data treatment that has been done. We have geo-tagged items with the countries quoted in the text. To do it, we built a word dictionary which allows to automatically detect countries.
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	Character	ISO3 code of countries detected in each item.
Nb_tag_detected	Numeric	Number of dictionary words which allowed detection of the country quoted in the item.

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

ID	Feed	time	Text1	Text2	TAG	Nb_tag_detected
3117022	en_AUS_austra_int	2014-10-01 00:33:26	Hong Kong protesters vow to stay put	HONG KONG demonstrators have rejected demands to end rallies that have paralysed the city ahead of a national holiday.	**HKG**	**1**
3117502	en_AUS_austra_int	2014-10-01 01:33:27	UK joins air strikes against jihadists	BRITISH warplanes have carried out their first raids against Islamic State, as the US launched multiple strikes against the jihadists in Iraq and Syria.	**GBR**	**2**
3117502	en_AUS_austra_int	2014-10-01 01:33:27	UK joins air strikes against jihadists	BRITISH warplanes have carried out their first raids against Islamic State, as the US launched multiple strikes against the jihadists in Iraq and Syria.	**IRQ**	**1**
3117502	en_AUS_austra_int	2014-10-01 01:33:27	UK joins air strikes against jihadists	BRITISH warplanes have carried out their first raids against Islamic State, as the US launched multiple strikes against the jihadists in Iraq and Syria.	**SYR**	**1**
3117502	en_AUS_austra_int	2014-10-01 01:33:27	UK joins air strikes against jihadists	BRITISH warplanes have carried out their first raids against Islamic State, as the US launched multiple strikes against the jihadists in Iraq and Syria.	**USA**	**1**
3117503	en_AUS_austra_int	2014-10-01 01:33:27	Secret Service lashed for Obama breach	THE head of the US Secret Service says an intrusion by a knife-wielding man at the White House was “unacceptable” and won’t occur again.	**USA**	**3**

The word dictionary used to geo-tag data is also provided (Dico_Country_Free.csv).

d. statistic report (html)

With the data, we also provide a statistic report (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 “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**