Visualizing Our Global World

Comprehensive Insights on Global News Imagery

Samvat Rastogi, Nirad Parhi, Sneha Lakshmi Nyayapati, Niketh Shetty, Sreya Chakrabarti

April 24, 2017

Abstract

The Global Database of Events, Language, and Tone (GDELT) is described as “an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what’s happening around the world, what its context is and who’s involved, and how the world is feeling about it, every single day”[7]. It was created by Kalev Leetaru and Georgetown University. In this project, we attempt to leverage the huge amount of information contained about the history of the world in the data sets of the GDELT project and use them to create insightful visualizations.

1 Introduction

The project aims to make meaningful analysis of the vast dataset provided by GDELT Project. The Global Database of Events, Language and Tone is one of the largest and the most comprehensive open database. It comprises data from the year 1979 till present consisting of a collection of news articles from across the globe. The primary source of data is the news media - print, broadcast and the web. GDELT has recorded data stored in three forms of datasets

1. Events database
2. Global Knowledge Graph (GKG)
3. Visual Global Knowledge Graph (VGKG)

The Events database has details about all physical activities happening around the globe.
The GKG converts all the raw data into a list of persons, geolocations, organizations etc. For the project we used the VGKG datasets. It was accessed using Google BigQuery.
The VGKG is an enhancement on top of the GKG, which uses the Google Cloud Vision API to interpret global news media. VGKG is powered to extract meaningful information from a stream of visuals that are part of the world’s news.
The visualizations shown in this paper intends to displayimportant information extracted and analysed from the VGKG dataset using statistical analysis.
The major questions we targetted to get address are:

- How are the major news themes varying in popularity, each quarter, over a particular time period?
- What are the major causalities of joy, sorrow, anger and surprise in the images tagged with each news article?
- What are the top geological sources for articles tagged as Medical, Violent, Adult and Spoof by the Google SafeSearch algorithm?
- What are the top labels occurring in news articles based on the frequency of their occurrence?

1.1 Burst Analysis on Labels

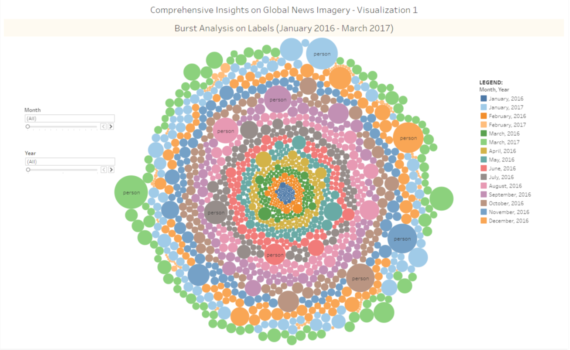
This Visualization intends to highlight the variation of interests in major news topics over time. For a time period ranging from January 2016 to March 2017, the major topics of interests represented by news articles are shown with partitions of monthly time slice. We have performed a Burst Analysis for this visualization.Each news topic is represented by a bubble. The size of the bubble is denoted by the number of news articles for the period in question.
We captured 15 such subsets of data from BigQuery for each month. Below is the sample query for extracting the Top 100 labels ranging from 1st January 2016 to 31st March 2017.
Column Used : Labels
Query Used :

```
select Description, count(1) as cnt from (
SELECT Date, REGEXP_REPLACE(SPLIT(Labels,
'<RECORD>'),r'<FIELD>.*', '') Description
FROM [gdelt-bq:gdeltv2.cloudvision]
where Labels is not null
and Date > 201703010000 and
Date < 20170331000000
) group by Description
order by cnt desc limit 100;
```

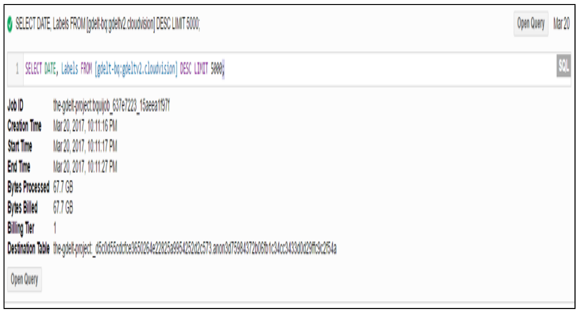
Amount of data processed : 69GB
Time Taken: 21 seconds
Rows chosen : Top 100
Sample Dataset:
One of the sample datasets , after all the preprocessing looks as below

Description	**cnt**
person	63574313
profession	32677970
vehicle	24395994
sports	16571878
speech	16258180
people	12822962
font	12571026
brand	11551766

Snapshot of the visualization :



Detailed Analysis
Below is the process of extracting and analyzing the data using Python, and using the resultant dataset to produce a Bubble Chart visualization in Tableau (Sample Visualization 1 in this document)
To get the gist of steps involved in data analysis for bubble chart (burst analysis), we took a sample of 5000 rows from data from the CloudVision table using below BigQuery, with its job details:



SELECT DATE, Labels FROM [gdelt-bq:gdeltv2.cloudvision] DESC LIMIT 5000;

```
SELECT DATE, Labels FROM [gdelt-bq:gdeltv2.cloudvision] DESC LIMIT 5000;
```

Job ID	the-gdelt-project:bquijob_637e7223_15aeea1f97f
Creation Time	Mar 20, 2017, 10:11:16 PM
Start Time	Mar 20, 2017, 10:11:17 PM
End Time	Mar 20, 2017, 10:11:27 PM
Bytes Processed	67.7 GB
Bytes Billed	67.7 GB
Billing Tier	1
Destination Table	the-gdelt-project:_d5c0d55cdcfce3650264e22825a9954252d2c573.anon3d75984372b06fb1c34cc3433d0d29ffc9c2f54a

As data is huge, the BigQuery (in SQL format) took 10 seconds to process the 67.7 GB of data. Further, a CSV was exported from the web console of BigQuery and to perform the next step of data analysis, we used Anaconda 3.5 (python distribution) in Jupyter Notebook. This notebook is available at this page with complete code and outputs.
The overview of analysis is as below:

1. Loaded csv into dataframe using pandas

package.

2. Checked for duplicates.
3. Removed the duplicates and removed rows with Labels = NULL
4. Processed the Labels column and separated each record (<RECORD>).
5. Then each of these records are again processed to extract label description.
6. Further, we checked the number of total unique timestamps in the data sample and found that there are total 11 unique time stamps with 4 of 2016 and 7 of 2017.
7. Extraction of year from the timestamp was done.
8. The command “labels.describe()” was used to check the summary of table to check the correctness of the process.
9. Finally, the size of the data, which was grouped by ‘DATE’ and ‘Labels’ respectively, was taken and following output (sample) was obtained.
10. This output was then exported into CSV and loaded into Tableau to create the Bubble Chart visualization.

The Bubble Chart visualizations is present here.
The page is interactive. The user can view the bubble charts for any given month from the year 2016 till date by choosing the appropriate year and month by using the provided sliders.

Insights gained from the visualization:

1. The label "person" is most frequently occurring in all the news articles across the span of 2 years.
2. The reason is pretty intuitive. The Cloud vision API easily identifies persons present in the images in the new articles.
3. This also indicates that persons are a generally occurring theme in new article images.
4. Other frequently occurring labels are - people, profession, vehicle, athlete, sports
5. We can conclude that the labels relating people like vehicle, sports are occurring most frequently as the day-to-day news articles are dominated by such related content.

1.2 Quantified Relationship between Emotions and Associated Labels

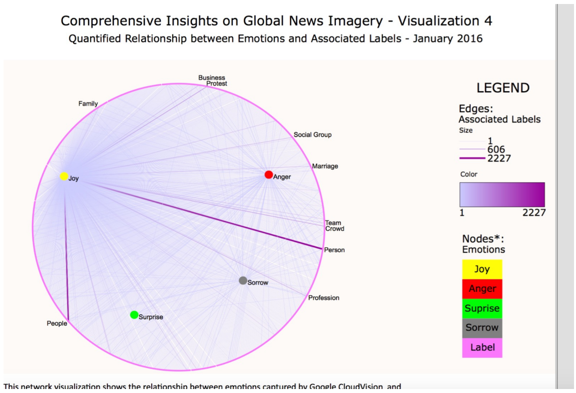
This visualization helps give some insight into the causes of **joy**, **sorrow**, **anger** and **surprise** by analyzing the likelihood of each emotion in the images tagged to the news articles in the dataset. The emotions are then compared to the label annotations describing the contents of the image.
The visualization is a network with each emotion as a source with the child nodes of each source denoting the causes of the emotion. The reason for choosing a network visualization for this use case is that a single label may be associated with multiple emotions. For example , a label “Person” is very commonly found and is seen to associate with all the 4 emotions listed. Hence a network comes across as the most relevant visualization , where the relationship between emotions and labels is represented by joining their nodes and their quantitative relation is shown by the thickness of the edges.

Columns Used: Faces,Labels
Sample Query:

```
Select count(*) as Count from (
SELECT Labels,
INTEGER(REGEXP_EXTRACT(SafeSearch)
r'^.*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>.*?<FIELD>.*?<FIELD>(.*?)<FIELD>.
*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>.*?<FIELD>.*?$')) sorrow,
INTEGER(REGEXP_EXTRACT(SPLIT(Faces,'<RECORD>'),
r'^.*?<FIELD>.*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>.*?<FIELD>.*?<FIELD>(.*?)<FIELD>.
*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>.*?$’)) anger,
INTEGER(REGEXP_EXTRACT(SPLIT(Faces,’<RECORD>’),
r’^.*?<FIELD>.*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>.*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>(.*?)<FIELD>.*?<FIELD>.*?<FIELD>.*?$’)) joy,
INTEGER(REGEXP_EXTRACT(SPLIT(Faces,’<RECORD>’),
r’^.*?<FIELD>.*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>.*?<FIELD>.*?<FIELD>.*?<FIELD>.*?<FIELD>.
*?<FIELD>(.*?)<FIELD>.*?<FIELD>.*?$’)) surprise,
FROM [gdelt-bq:gdeltv2.cloudvision]
where Faces is not null
and Date > 201601010000 and Date < 20170430000000
) where sorrow > 0 or anger > 0
 or joy > 0 or surprise > 0 ;
```

Data Processed : 34.5 GB

Snapshot of the visualization:



This network visualization shows the relationship between emotions captured by Google CloudVision, and

Insights gained from the visualization:

1. Out of the four categories of emotions i.e. "Joy", "Sorrow", "Surprise" and "Anger" , the emotion "Joy" is most frequently occuring.
2. The top labels associated with all the 4 emotions are - Person, People, Protest, Crowd, Business, Social Group, Marriage etc.
3. The labels occurring most frequently with "Joy" are Family, Person, People, Profession etc.
4. Similarly, we can see other frequently occurring labels with emotions.

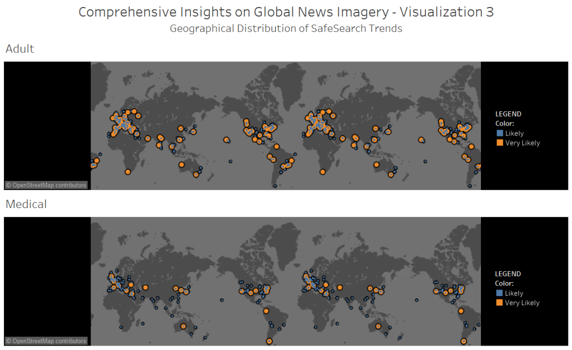
Detailed Analysis:
There is a lot of data processing that went into this visualization. The first step was to filter out each label from a given row using the REGEXP function in SQL. The next step was to find their corresponding emotions. Emotions with values greater than 0 have been chosen,since they represent the likeliness of the emotion detected by the API. The step by step analysis for this visualization using Python is present here. The final visualization can be found here.

1.3 Geographical distribution of SafeSearch trends

This visualization depicts the sources of news articles categorized as **Medical**, **Violent**, **Adult** and **Spoof** by the Google SafeSearch algorithm on a two-dimensional map of the Earth. The Geolocations are marked with the number of articles in each category originating from the location in question.
For this visualization, the two fields in use are **GeoLandmarks** and **SafeSearch**. We took the subset of the data which has proper values for both the fields for the entire date range. The Geolandmarks column is highly sparse hence , we have filtered out the data and considered only the ones with Not Null values , which comes down to 3778206 rows.
Sample Query:

```
select GeoLandmarks, SafeSearch
from [gdelt-bq:gdeltv2.cloudvision]
where GeoLandmarks is not null
and SafeSearch is not null
and DATE > 20160101000000
and DATE < 20170430000000;
```

Amount of data processed : 8.89 GB
Time Taken : 27 seconds
Snapshot of the visualization :


Comprehensive Insights on Global News Imagery - Visualization 3
Geographical Distribution of SafeSearch Trends
Adult
© OpenStreetMap contributors
LEGEND
Color:
Likely
Very Likely
Medical
© OpenStreetMap contributors
LEGEND
Color:
Likely
Very Likely

Detailed Analysis :

This data was again processed in Tableau by using the 'Custom Split' feature. This feature enables us to split the values from the SafeSearch field into individual columns , so that mapping could be done with respect to each category i.e. Violence , Medical , Spoof and Adult. Also, the attributes "Latitude" and "Longitude" were merged in each row of the data separated by a comma as the delimiter. Initial split for the GeoLandmarks was done on the basis of the <RECORD>and <FIELD>delimiters , and later the latitude and longitude were retrieved by splitting it on a comma. Sample GeoLandmarks row looks as below :

White House<FIELD>0.32624927
<FIELD>/m/0278psl
<FIELD>38.897312,
-77.036564<FIELD><RECORD>White House
<FIELD>0.18719363
<FIELD>/m/081sq
<FIELD>38.902255, -77.037162<FIELD >

The highlighted values are the latitude and longitude. Hence there was a lot of cleaning and pre-processing that had to be done to obtain these values. After the data was ready , we plotted these categorical values against each location on the World Map using Tableau. It is shown as a dashboard view, and the current output of the visualization can be found at this link.
Each category has been represented for **'Likely'** and **'Highly Likely'** values. The size and the color coding stand for the same. The legend is mentioned on the same Dashboard.

Insights gained from the visualization :

1. Adult content in the news articles are majorly generated from North America and Europe Regions.
2. In case of medical content in news, Europe region dominates the U.S..South Africa and Australia also make news in the medical field.
3. News articles coming under spoof category is generated in almost equal proportion from the developed countries.
4. News articles containing violent images are more predominantly generated from Europe, Russia, India and U.S.

1.4 Top Labels based on SafeSearch category

This visualization is a TreeMap portraying the top 20 labels for each category of the Safe Search field. Through this visualization we are trying to find the most common labels or objects present in an image that link an article to the type of search.
The type of search can be found out from the values of the column 'SafeSearch'.
A sample value of this column looks like '2<FIELD>2<FIELD>2<FIELD>2'.
There are four categories of search that can be represented here – **Violent** , **Medical** , **Spoof** and **Adult**.
In the above example every category has a value of 2. The types of values these categories can take are as below :

- -2 : Very Unlikely
- -1 : Unlikely
- 0 : Undecided
- 1 : Likely
- 2 : Very Likely

In our visualization we are trying to show the labels that are 'Likely' or 'Highly Likely' to belong to a certain category.
For example , the top 20 labels or objects that decide that the image has violent content can be found out using the below query :

```
select Label ,count(1) as cnt,SafeSearch from
(SELECT SafeSearch ,REGEXP_REPLACE(SPLIT(Labels,
'<RECORD>'),r'<FIELD>.*', '') Label
FROM [gdelt-bq:gdeltv2.cloudvision]
where Labels is not null
)where SafeSearch like '1<FIELD>\%'
or  SafeSearch like '2<FIELD>\%'
group by Label,SafeSearch
order by cnt desc limit 50;
```

Queries used for other categories are listed below: Medical Searches:

```
select Label ,count(1) as cnt,SafeSearch from
(SELECT SafeSearch ,REGEXP_REPLACE(SPLIT(Labels,
'<RECORD>'), r'<FIELD>.*', '') Label
FROM [gdelt-bq:gdeltv2.cloudvision]
where Labels is not null)
where SafeSearch like '_<FIELD>1\%'
```

```
or  SafeSearch like ’_<FIELD>2\%’
group by Label,SafeSearch
order by cnt desc limit 50;
```

Spoof:

```
select Label ,count(1) as cnt,SafeSearch from
(SELECT SafeSearch ,REGEXP_REPLACE(SPLIT(Labels,
’<RECORD>’), r’<FIELD>.*’, ’’) Label
FROM [gdelt-bq:gdeltv2.cloudvision]
where Labels is not null
)where SafeSearch like ’_<FIELD>_<FIELD>2\%’
or  SafeSearch like ’_<FIELD>_<FIELD>1\%’
group by Label,SafeSearch order by cnt desc limit 50;
```

Adult :

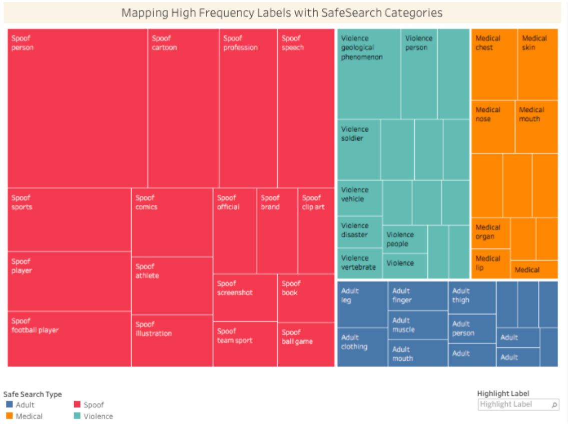
```
select Label ,count(1) as cnt,SafeSearch from
(SELECT SafeSearch ,REGEXP_REPLACE(SPLIT(Labels,
’<RECORD>’), r’<FIELD>.*’, ’’) Label
FROM [gdelt-bq:gdeltv2.cloudvision]
where Labels is not null
)where SafeSearch like ’_<FIELD>_<FIELD>_<FIELD>1\%’
or  SafeSearch like ’_<FIELD>_<FIELD>_<FIELD>2\%’
group by Label,SafeSearch order by cnt desc limit 50;
```

Data Processed : 73.8 GB for each query
Time Taken : 12 seconds for each case

Sample dataset looks like:

Label	cnt	SafeSearch
leg	3017	0<FIELD>0<FIELD>0<FIELD>2
person	2987	0<FIELD>0<FIELD>0<FIELD>2
clothing	2639	0<FIELD>0<FIELD>0<FIELD>2
black hair	2392	0<FIELD>1<FIELD>0<FIELD>2
finger	2285	0<FIELD><FIELD>0<FIELD>2
chest	2143	0<FIELD>1<FIELD>0<FIELD>2
muscle	2127	0<FIEL>1<FIELD>0<FIELD>2
leg	2094	0<FIELD>1<FIELD>0<FIELD>2
hand	2053	0<FIELD>1<FIELD>0<FIELD>2
mouth	1945	0<FIELD>1<FIELD>0<FIELD>2

Snapshot of the visualization:



Insights gained from the visualization :

1. The top labels which generally trigger the Spoof category of Cloud Vision API are Person, Cartoon, profession, speech, sports etc.
2. These are the most common labels that occur in news article which are not normally related to definite categories like violence etc.
3. For violence SafeSearch category, the top labels are - geological phenomenon, soldier, vehicle, disaster etc.
4. Similarly, for the category "Medical", top labels are - chest, skin, nose, mouth, human body.

The final visualization can be found at link.

2 Discussion of related work

After doing a detailed search about the visualization we plan to showcase, we did not find any similar projects. There were projects which discussed daily trends of conflict across the globe [3]. This project tries to find and correlate specific news items related to conflicts, analyze them and generate various reports summarizing them. Another project [4] tries to find geographical details from news articles and geocodes those articles to specific locations. With respect to sentiment analysis, there have been projects like [5]. This project tries to portraythe positive or negative tone of the news articles based on the geolocation. It finally depicts the average emotion of various parts of the globe on a flat map. There is also an existing Word Cloud visualization over the top entries of a GDELT GKG field for a specific search. The sample can be seen at [6]. However, we are planning to depict a Word Cloud of the top 100 labels from the VGKG dataset and color code it as per the categories of the ‘Safe Search’ field.

Our project consists of four visualizations and the ideas of portraying each one has been chosen as per the relation/mapping between the relevant fields. There have been similar concepts over the web, but we haven’t found anything similar to our envisioned work.

3 Statistics of the data sets used

The dataset used in this project is the VGKG dataset present in the CloudVision table in BigQuery. The table is 2.10 TB in size and has 12 columns, few of which are Date, DocumentIdentifier, ImageURL, Labels, Faces etc. It has a total of 248,841,467 records and has data from Dec 31,2015 to the present. We shall be using the below columns and their corresponding attributes as listed below, for extracting relevant data for the proposed visualizations:

- **Date** – It indicates the date when a given image was last monitored by the Cloud Vision API.
- **Labels** - We shall be mining in the ’description’ part of this entity. This attribute alone will help us with three of our visualizations - Finding causes behind emotions, Word Cloud representation of top 100 labels with their corresponding SafeSearch category, and Finding the major headline making categories out of all these news articles.
- **Faces** - Here we shall be extensively using four attributes - EmotionSorrowLikelihood,EmotionAngerLikelihood, EmotionJoyLikelihood, EmotionSurpriseLikelihood.
- **GeoLandmarks** - The attributes ’Latitude’ and ’Longitude’ shall be put into use for visualizing the type of content coming from certain parts of the world. This shall be represented over a flat map.
- **SafeSearch** - We shall be extracting information based on the following attributes -ViolenceLikelihood, MedicalLikelihood, SpoofLikelihood, AdultLikelihood.

4 Problems surfaced during our work

- As the dataset that we are working on is in the scale of Tera Bytes, the output of the queries that we fire on Google BigQuery sometimes ranges in GBs. In such cases, as there is a quota (in bytes) allocated to each user, we are running into errors and exceptions. We are not able to get a view of the resultant datasets in such cases.
- Since the existing data is in the form of JSON, we might have to do a lot of pre-processing and cleaning in order to make it usable for tools like Sci2.

5 Challenges faced during the project.

Some of the challenges that we have faced till now in the project are:

- Only around 1.5 percent of the dataset has geographic information. The rest of the dataset does not have geolocation specific details. Hence, we have limited data set to show geographic trends and patterns which could have resulted in insightful visualizations.
- Data is sparse in some of the columns of the dataset hence affecting the quality of data and the scope of finding strong correlations.
- The volume of the dataset has been a concern so far in the project. It restricts us to use only the Google BigQuery API for mining, extraction and analysis.
- The Google CloudVision API parses the images in the news articles and identifies them as common nouns like person,people,athlete etc.It does not recognize proper nouns such as name of a person or any particular historical event. With such data , it is a challenge to find insightful/historical information to be presented.
- The data set contains data ranging from 2016 - 2017. Hence it poses a limitation in performing any kind of detailed temporal analysis.

6 Problems surfaced during validation and resolution :

The biggest challenge we came across for this project was dealing with the data. With data in the scale of Terabytes, it was difficlut to access the entire dataset for any given criteria. Fine tuning the queries helped us to extract the relevant data .

Sparsity of data was another issue , since finding a co-relation became difficult due to the huge number of Null values in various columns. We handled this scenario by filtering the data which has significant values in each column.

Since the dataset is very broad in terms of areas covered, it becomes difficult to portray a Network type of visualization.

References

[1] Börner, K. 2015. Atlas of Knowledge: Anyone Can Map. Cambridge, Massachusetts: The MIT Press.

[2] http://blog.gdeltproject.org/announcing-the-new-gdelt-visual-global-knowledge-graph-vgkg

[3] http://blog.gdeltproject.org/gdelt-daily-trend-reports/

[4] https://www.forbes.com/sites/kalevleetaru/2017/02/21/visual-geocoding-a-quarter-billion-global-news-photographs-using-googles-deep-learning-api/#68bf7a1217fa

[5] https://www.forbes.com/sites/kalevleetaru/2017/02/22/mapping-global-happiness-in-2016-through-a-quarter-billion-news-articles/#4141a7642692

[6] http://analysis.gdeltproject.org/module-gkg-wordcloud.html

[7] http://gdeltproject.org/about.html