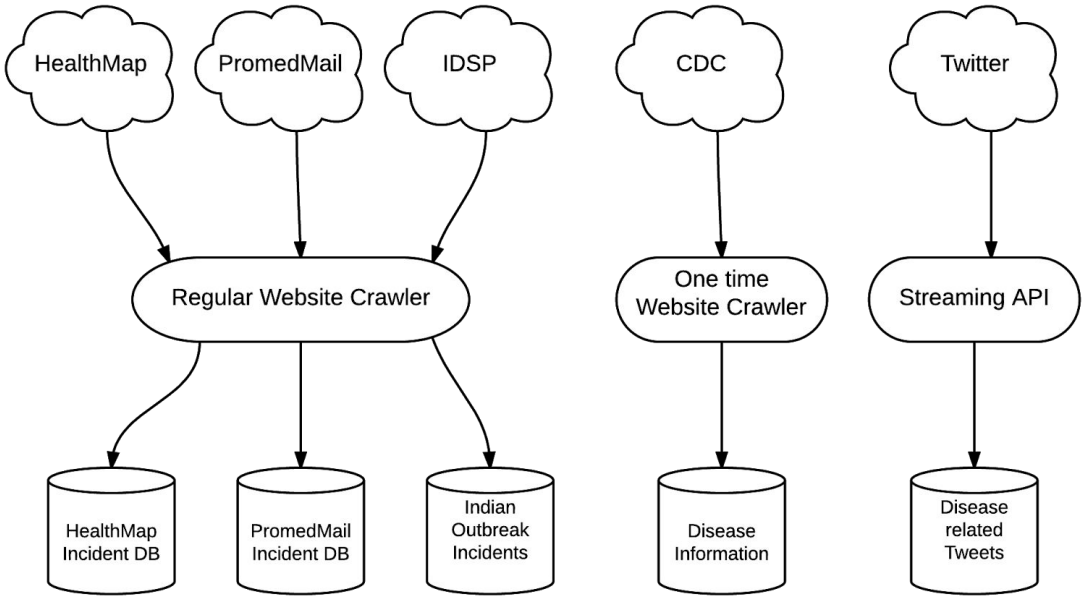
Our System can be explained with following Diagram -:



In this system , we have set up neverending website crawlers on HealthMap , PromedMail and Integrated Disease Surveillance Programme (IDSP) which runs at regular interval and update our HealthMap Incident DB, PromedMail Incident DB and Indian Outbreak Incidents databases respectively with any new information regarding the incidents available .

Similarly One Time Crawler was set up on Centers for Disease Control and Prevention (CDC) website to collect information regarding the diseases that have been discovered till now in Disease Information DB.Keyword monitoring is performed on Twitter using Streaming API to get the Disease related Tweets.

We have developed a system that crawls and store information related to diseases from following **sources**:

1. **HealthMap** : (http://www.healthmap.org/en/)

This source of information collects the outbreak information from Google News , some influential twitter handlers and Promed Mail . Healthmap currently tries to identify location , number of people affected / died ,disease name and species affected from such outbreak reports/news(above mentioned sources).

We have collected around **539,919** such incident report around the globe from **August 1994 to September 2015** . This data keeps on updating itself every hour for any such incident reported on HealthMap.

Such data from Promed Mail and Google News also provide us with the relevant news headlines related to each incident as reported manually (in case of Promed-Mail) or in Google News . These sources of new headlines are also being stored by us and added to list of news sources that can be monitored daily for any information of such incidents .

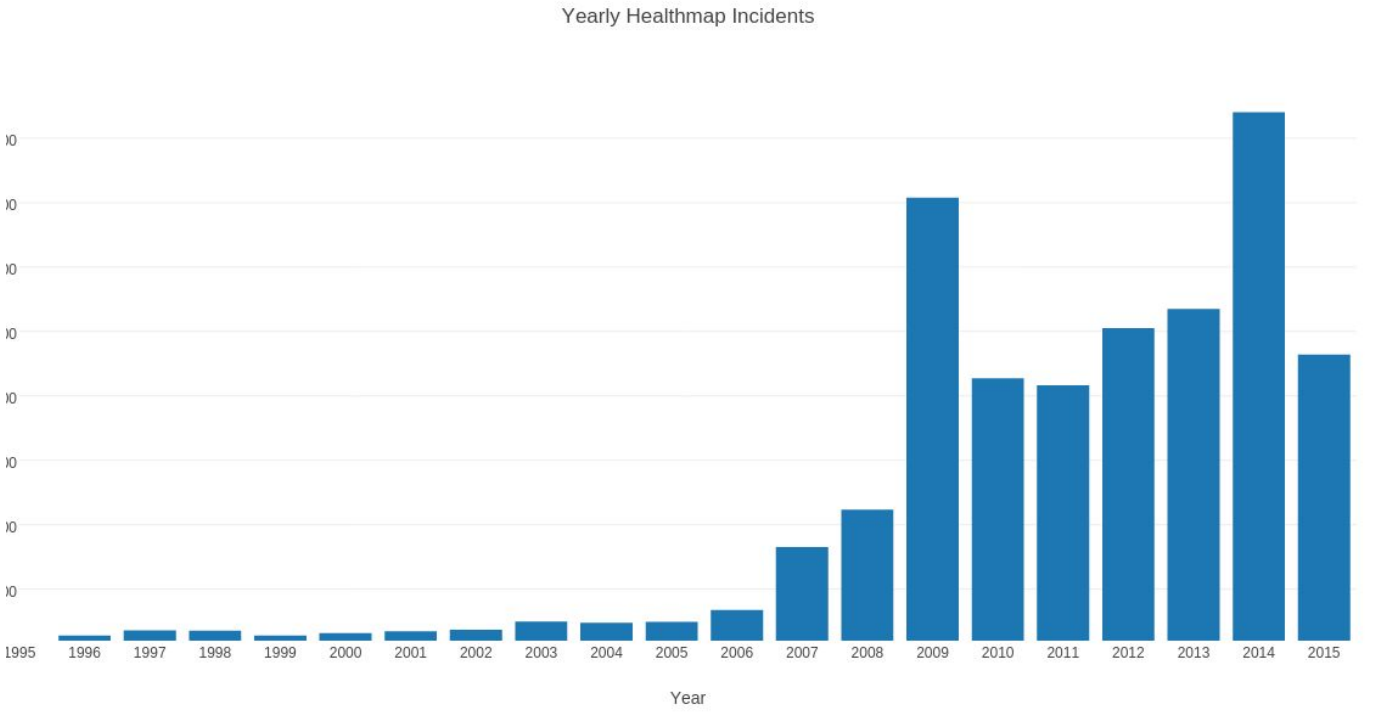
This data can either be used to validate/enhance the information we get from twitter or it can also be used to compare our results .

Structure of HealthMap DB is given as -:

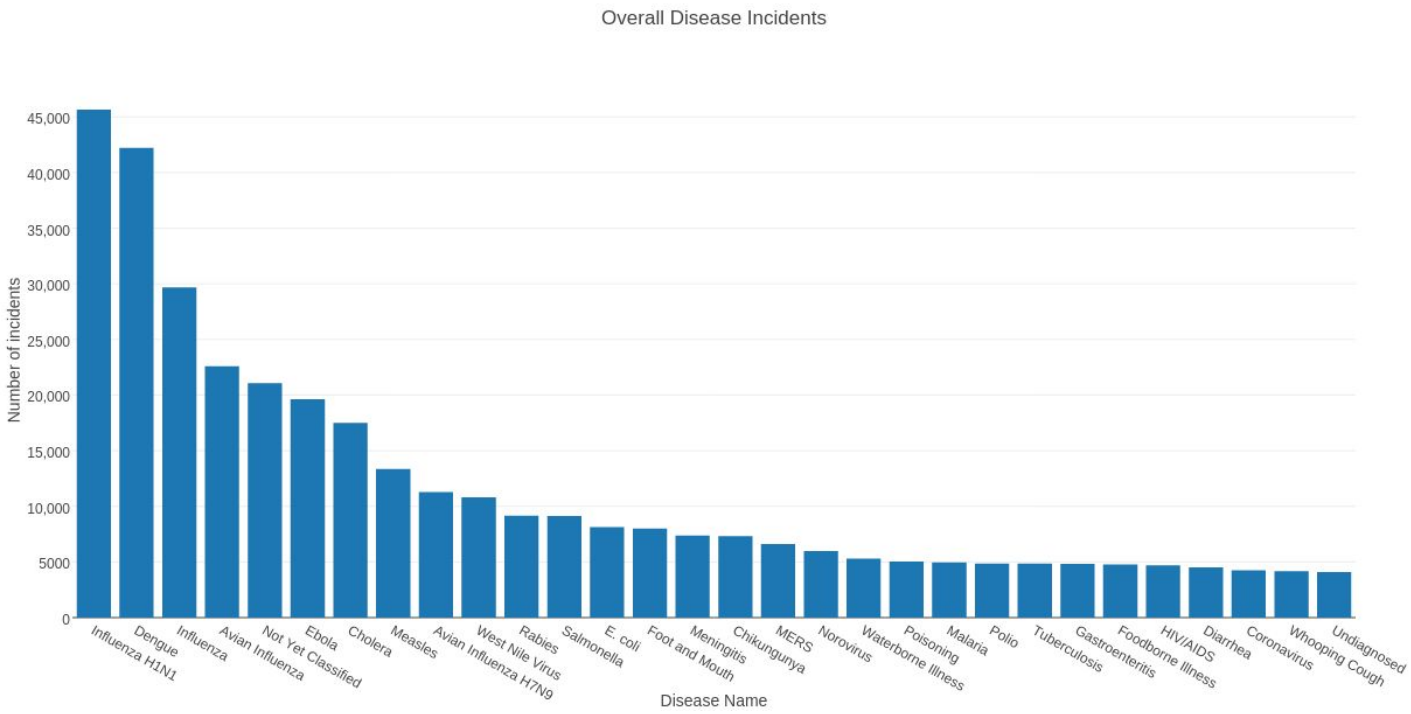
```
CREATE TABLE IF NOT EXISTS `Disease_HealthMap_Full` (
  `Date` date NOT NULL,
  `Source` varchar(100) NOT NULL,
  `Disease` varchar(100) NOT NULL,
  `Location` varchar(150) NOT NULL,
  `Lattitude` varchar(150) NOT NULL,
  `Headline_Report` text NOT NULL,
  `Summary` text NOT NULL,
  `Article_URL` varchar(250) NOT NULL,
  `Species` varchar(100) NOT NULL,
  `Cases` int(11) DEFAULT NULL,
  `Death` int(11) DEFAULT NULL,
  `Rating` int(11) NOT NULL,
  `ID` int(11) NOT NULL,
  UNIQUE KEY `Date` (`Date`,`Source`,`Disease`,`Location`,`ID`,`Article_URL`,`Species`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

Sample Values -:

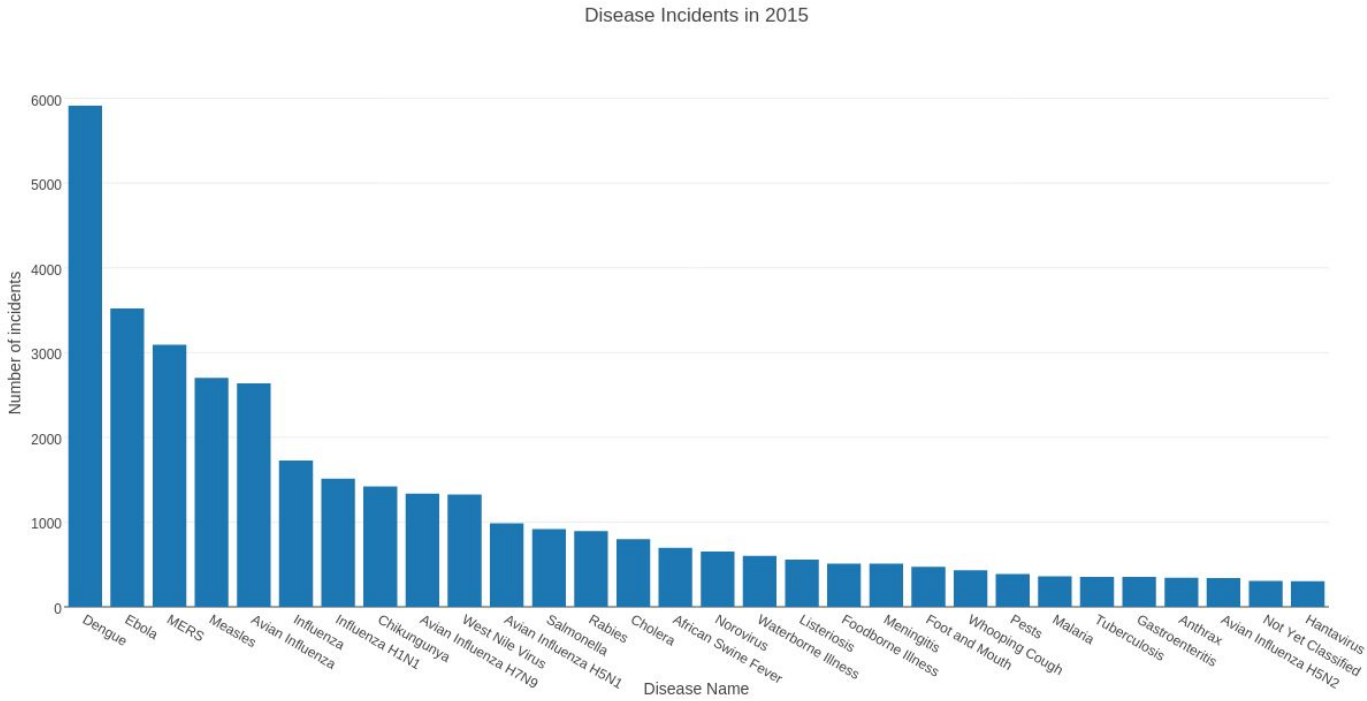
Date	Source	Disease	Location	Lattitude	Headline_Report	Summary	Article_URL	Species	Cases	Death	Rating	ID
2015-08-23	Google News	Not Yet Classified	Florida, United States	'28.149500','-81.650398',5,212	Machine could help map citrus drop patterns - The Ledger	'Machine could help map citrus drop patterns The fatal bacterial disease citrus greening was confirmed in Florida around 2005. It wasn't until the 2012-13 season, however, that the disease's spread and intensity began showing up in extraordinary levels of mature oranges, grapefruit and tangerines ;... ', 'Florida, United States	Not Yet Classified	Humans'	http://www.theledger.com/article/20150823/NEWS/150829853	Humans	*NULL*	*NULL*	1	3596514
2015-08-23	Google News	Ebola	Nigeria	'9.593960','8.105310',5,62	Coming Soonu202693 Days, A Film on Ebola - THISDAY Live	u'Coming Soonxe2x80xa693 Days, A Film on Ebola Behind-the-scene activities for the production of the movie, 93 Days: A Nation's Story on Survival and Sacrifice, are revving up a notch. The movie, based on the true-life story of men and women who risked their lives and made sacrifices to save their ;... ', 'Nigeria	Ebola	Humans'	http://www.thisdaylive.com/articles/coming-soon-93-days-a-film-on-ebola/218219/	Humans	*NULL*	*NULL*	4	3596555



Analysis of Entire HealthMap DB



Analysis of HealthMap DB for Incidents in Year 2015



2. Centers for Disease Control and Prevention , **CDC** (http://www.cdc.gov/)

This source of information is used to get the name of all possible diseases , their categories , causing agent as well as various abbreviation and common names for each disease . This source is crawled daily to get information related to any new diseases updated by CDC .

Using this crawler , we have collected information for around **814** diseases along with their type

Structure of Disease Information is given as -:

```
CREATE TABLE IF NOT EXISTS `disease_name_cdc` (
  `DISEASE_NAME` varchar(100) NOT NULL,
  `TYPE` varchar(100) NOT NULL,
  `LINK` varchar(500) NOT NULL,
  `OTHER_NAMES` varchar(500) NOT NULL,
  `NUMBER_NAMES` int(11) NOT NULL,
  UNIQUE KEY `DISEASE_NAME` (`DISEASE_NAME`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

Sample Values -:

DISEASE_NAME	TYPE	LINK	OTHER_NAMES	NUMBER_NAMES
Abdominal Aortic Aneurysm	dhdsp	http://www.cdc.gov/dhdsp/data_statistics/fact_sheets/fs_aortic_aneurysm.htm	'Abdominal Aortic Aneurysm', 'Aortic Aneurysm', 'Aortic Dissection', 'Thoracic Aortic Aneurysm'	4
ACE	nccdphp	http://www.cdc.gov/nccdphp/ace/	'ACE', 'Adverse Childhood Experiences'	2
Acinetobacter Infection	HAI	http://www.cdc.gov/HAI/organisms/acinetobacter.html	'Acinetobacter Infection'	1
Acquired Immune Deficiency Syndrome	Disease	http://www.cdc.gov/hiv/	'Acquired Immune Deficiency Syndrome', 'Acquired Immunodeficiency Syndrome', 'AIDS', 'HIV/AIDS', 'Human Immunodeficiency Virus'	5

3. Program for Monitoring Emerging Disease , **PromedMail** (http://www.promedmail.org/)

This source of information consist of manually created reports by people throughout the world which is approved and published . These reports contain headlines and summary where headline generally indicate country of occurrence of disease and name of disease .Summary can be used to know more about the incident and also involves source to validate the report .

This website is crawled daily to keep our information updated . **44149** such incidents have been reported and collected by us till now after required preprocessing to remove irrelevant articles such as informative articles which do-not have any information related to outbreaks .

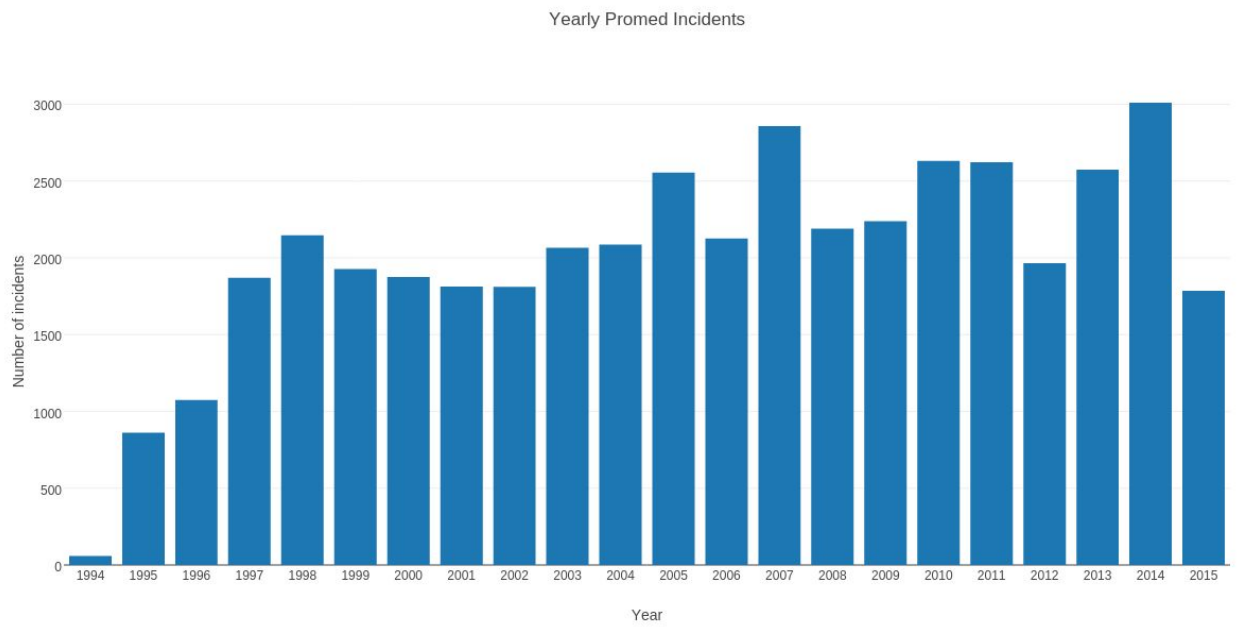
Structure of PromedMail DB is given as -:

```
CREATE TABLE IF NOT EXISTS `PROMED_DB` (
  `RELEASE_DATE` date NOT NULL,
  `DISEASE_NAME` varchar(100) DEFAULT NULL,
  `DISEASE_FULL` varchar(500) DEFAULT NULL,
  UNIQUE KEY `RELEASE_DATE` (`RELEASE_DATE`,`DISEASE_NAME`,`DISEASE_FULL`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

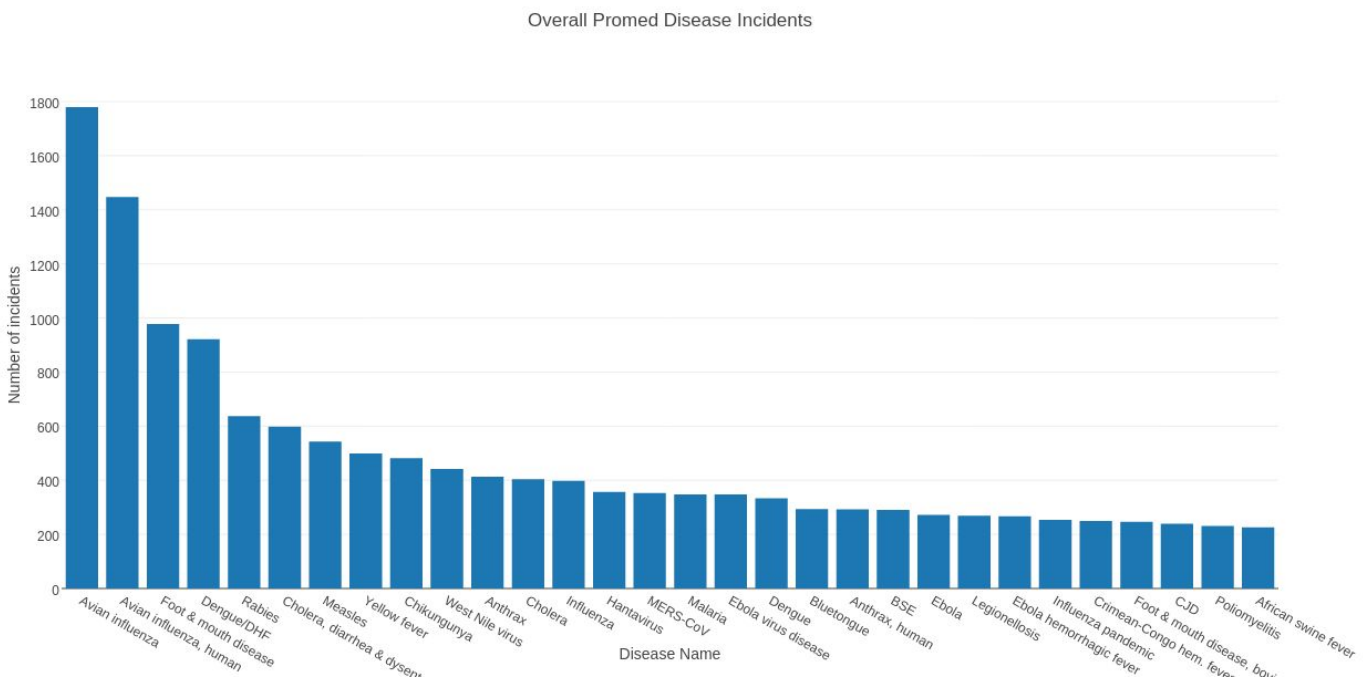
Sample Values -:

RELEASE_DATE	DISEASE_NAME	DISEASE_FULL
2015-09-01	Undiagnosed mortality, caprine	Undiagnosed mortality, caprine - India: (TG) RFI
2015-09-01	Poliomyelitis	Poliomyelitis: update (04): Ukraine, 2 cases, RFI
2015-09-01	Kerala wilt, coconut palm	Kerala wilt, coconut palm - India: (KL) susp
2015-08-31	Yellow Fever	Yellow Fever - Africa (05): Sudan, fake cards
2015-08-31	Sheath blight, rice	Sheath blight, rice - India: (PB)

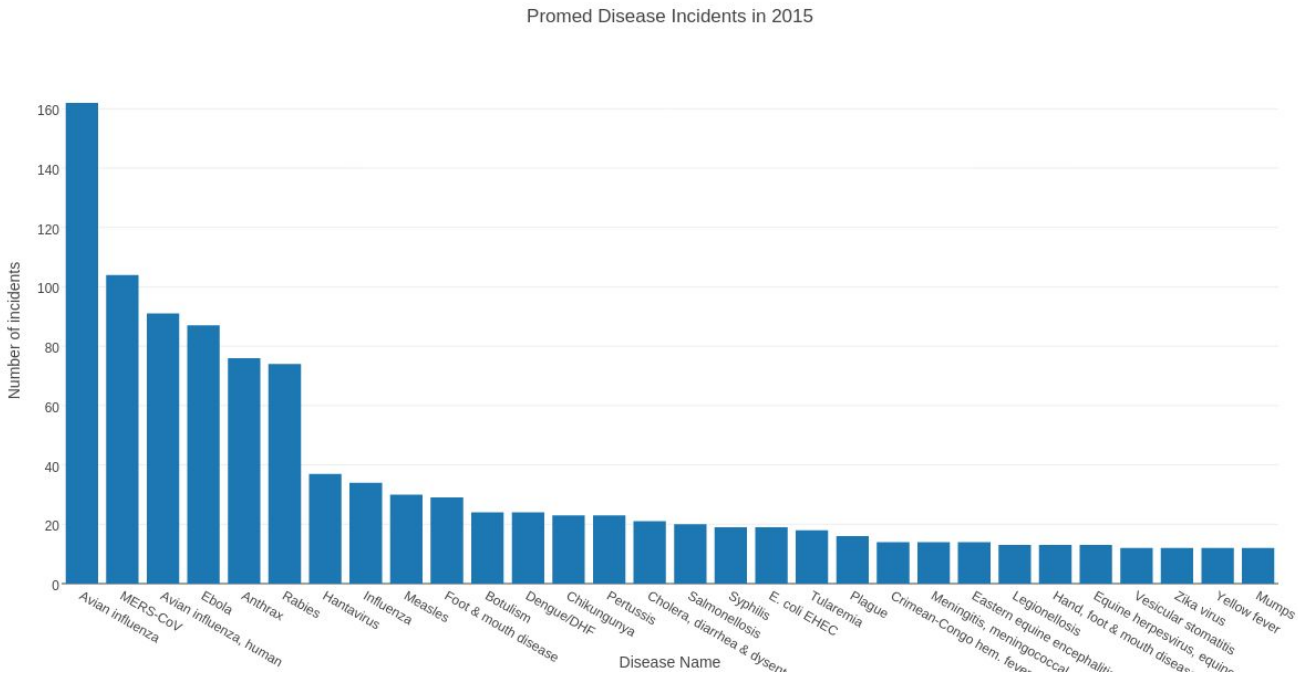
Overall Yearly Incidence Graph



Analysis of Entire PromedMail DB



Analysis of PromedMail DB for Incidents in Year 2015



4. Integrated Disease Surveillance Programme,**IDSP**(http://www.idsp.nic.in/idsp/IDSP/outbreaks.htm)

This source of information consist of all manually created PDF's for each month since 2009 that provides all the disease incidents reported in Indian states along with the cities in which they occur and number of people affected/died due to incident . The report also contain summary of each incident .

These are reports that can be used to check if system developed by us can be used for deep analysis of incidents within a particular country . As these incident reports are highly detailed and are crawled weekly to check if any new DF report is available for reports of next month .

Information related to **8,785** such incidents have been collected since **2009** .

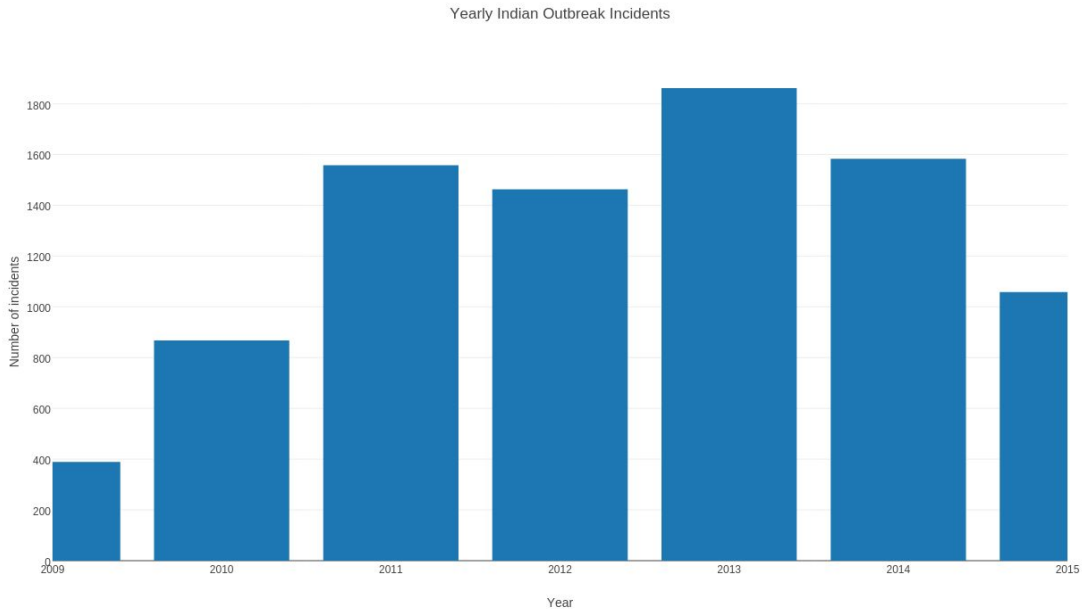
Structure of Indian Outbreak Incidence DB is given as -:

```
CREATE TABLE IF NOT EXISTS `Disease_incidents_India` (
  `City` varchar(100) NOT NULL,
  `Disease` varchar(100) NOT NULL,
  `Cases` int(11) NOT NULL,
  `Death` int(11) NOT NULL,
  `Start_Date` date NOT NULL,
  `Reporting_Date` date DEFAULT NULL,
  `Status` varchar(100) NOT NULL,
  `Filename` varchar(100) NOT NULL,
  `Report` text NOT NULL,
  UNIQUE KEY `Unique_key` (`City`,`Disease`,`Cases`,`Death`,`Start_Date`,`Status`,`Filename`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

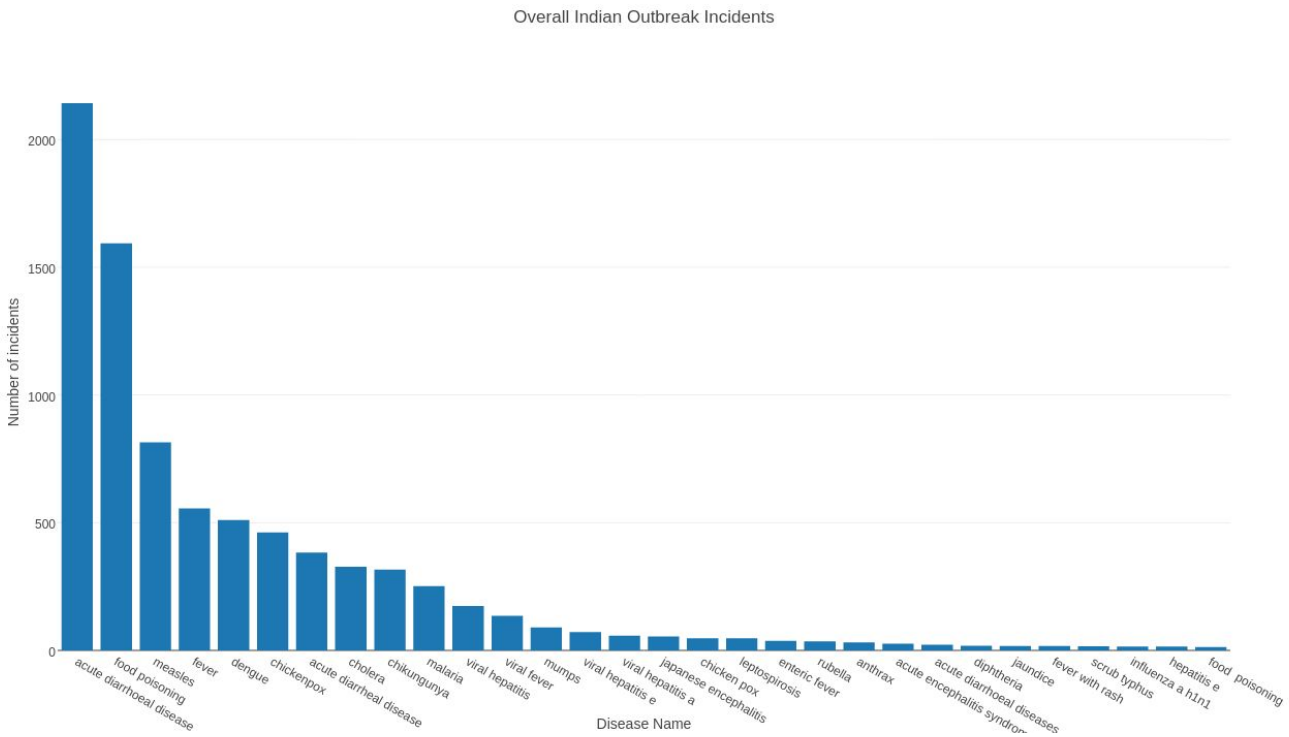
Sample Values -:

City	Disease	Cases	Death	Start_Date	Reporting_Date	Status	Filename	Report
thrissur	hepatitis a	10	0	2104-07-21	*NULL*	under surveillance	32nd_wk14	Cases reported from Pootharkkal, Block Cherpu, SC / Village Chevvur, District Thrissur. Index case was a patient with history of frequent travelling and eating outside food. He was treated for Hepatitis A in June. Other cases which developed jaundice were his friends and residing in neighborhood. They had history of consuming food together with index case
chandauli	chicken pox	12	0	2104-07-21	2014-08-07	under control	32nd_wk14	Cases of fever with rash reported from Village Ahikaura, SC Baheri, Block Dhanapur, District Chandauli. RRT investigated the outbreak. Active search for cases done. All cases belonged to 2 5 years age group. Symptomatic treatment given to cases. Health education given regarding isolation of cases & personal hygiene.
malappuram	measles	11	1	2015-12-26	*NULL*	under surveillance	4th_wk15	Cases of fever with rash reported from Villages Kuzhimanna, District Mallapuram. District RRT investigated the outbreak. House to house survey done.All Serum samples sent to NIV Bangalore tested positive for measles IgM ELISA. All the cases were isolated and treated symptomatically .Vitamin A was given to all the cases. Health education given.
delhi	measles	19	0	2015-12-16	*NULL*	under surveillance	3rd_wk15	Cases reported from Sarai Kale Khan and Batla house areas of District South Delhi. District RRT investigated the outbreak. Majority of cases were less than 06 yrs of age. Out of 10 samples tested at NCDC Delhi, 06 samples found to be positive for measles IgM ELISA. All cases treated symptomatically and provided with vitamin A. Health education given.

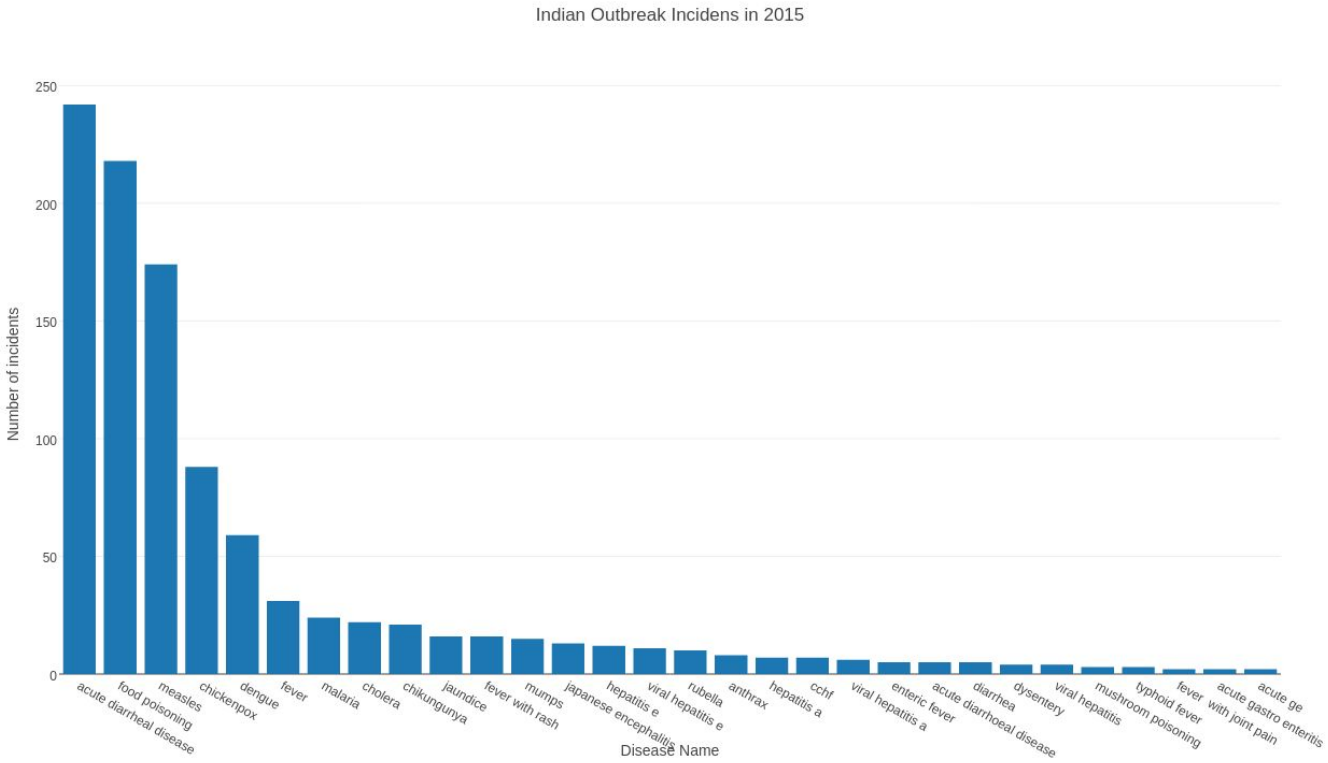
Overall Yearly Incidence Graph



Analysis of Entire Indian Outbreak Incidence DB



Analysis of Indian Outbreak Incidence DB for Incidents in Year 2015



5. **Twitter** (Using Specific Disease Name with Keyword such as "Ebola")

This source can be used for gathering information and views related to incidents /diseases . We have collected around **97493** multilingual tweets from around the world related to term "Ebola" since 27 Aug 2015. Similarly information related to twitter Users is also recorded in different table with foreign key relation to tweet table .

Ist Table :: Structure of Disease Related Tweet DB is given as -:

```
CREATE TABLE IF NOT EXISTS `Disease_Tweets` (
  `Created_at` datetime NOT NULL,
  `ID` varchar(100) NOT NULL,
  `Tweet` text NOT NULL,
  `Screen_Name` varchar(100) NOT NULL,
  `Urls` varchar(200) NOT NULL,
  `Expanded_Urls` text NOT NULL,
  `User_Mentions` varchar(150) NOT NULL,
  `Hashtags` varchar(100) NOT NULL,
  `Quoted_Tweet_ID` varchar(100) NOT NULL,
  `Retweet_ID` varchar(100) NOT NULL,
  `Lang` varchar(100) NOT NULL,
  PRIMARY KEY (`ID`),
  KEY `Screen_Name` (`Screen_Name`),
  KEY `Quoted_Tweet_ID` (`Quoted_Tweet_ID`),
  KEY `Retweet_ID` (`Retweet_ID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

Sample Values -:

Created_at	ID	Tweet	Screen_Name	Urls	Expanded_Urls	User_Mentions	Hashtags	Quoted_Tweet_ID	Retweet_ID	Lang
2015-09-03 18:37:33	639507540573847552	RT @blazednconfus3d: @MrsAnahata they talk about ebola in breaking bad too which aired in 2010!	MrsAnahata			blazednconfus3d , MrsAnahata			639499413870784512	en
2015-09-03 18:17:26	639502475486199808	RT @Holborncompany: All those of you who were asking what had happened to our old friend Mr Ebola, hang loose, there's a new kid in town ht…	AmbushPredator	https://t.co/fzh7tcXSc7	https://twitter.com/AJENews/status/639498487961374720	Holborncompany		639498487961374720	639499258933182466	en
2015-09-03 18:16:09	639502154789601281	RT @EbolaOutbreakUS: @shelliecorreia 40 United States Hospitals inject Americans with Ebola vaccine fever is COMMON side effect http://t.co…	trixienovel	http://t.co/cFU5DASEcl	http://www.ebolaoutbreakmap.com/listings/ebola-south-dakota-sanford-health-needs-30-people-to-be-injected-with-ebola-vaccine/	EbolaOutbreakUS , shelliecorreia			639499162803810304	en

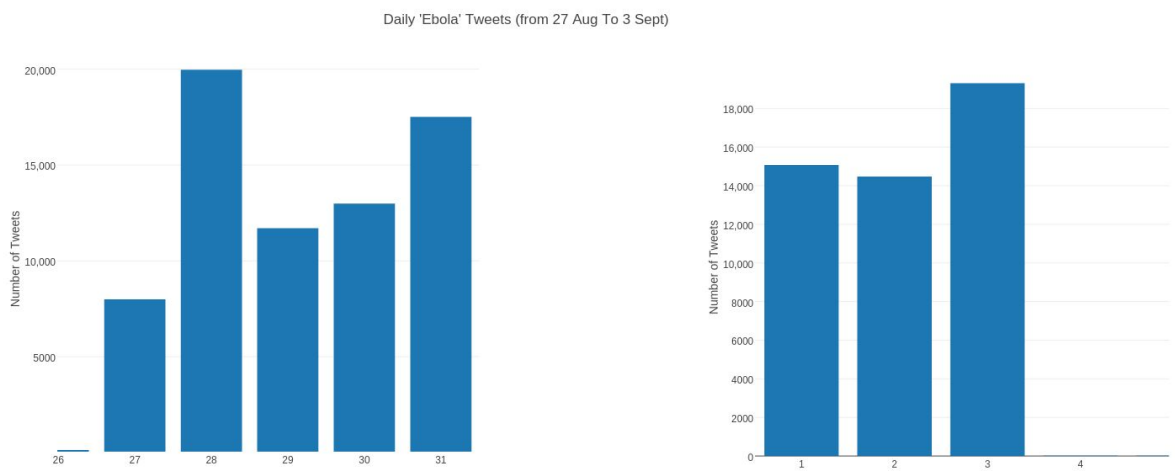
IInd Table :: Structure of Tweet UserName is given as -:

```
CREATE TABLE IF NOT EXISTS `Disease_Tweets_UserName` (
  `User_ID` varchar(100) NOT NULL,
  `Screen_Name` varchar(100) NOT NULL,
  `Name` varchar(100) NOT NULL,
  `Location` varchar(150) NOT NULL,
  `URL` varchar(200) NOT NULL,
  `Description` text NOT NULL,
  `Followers_Count` int(11) NOT NULL,
  `Friends_Count` int(11) NOT NULL,
  `Listed_Count` int(11) NOT NULL,
  `Favourites_Count` int(11) NOT NULL,
  `Statuses_Count` int(11) NOT NULL,
  `Created_At` datetime NOT NULL,
  `Time_Zone` varchar(150) NOT NULL,
  `Geo_Enabled` varchar(100) NOT NULL,
  `Lang` varchar(100) NOT NULL,
  PRIMARY KEY (`Screen_Name`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;
```

Sample Values -:

User_ID	Screen_Name	Name	Location	URL	Description	Followers_Count	Friends_Count	Listed_Count	Favourites_Count	Statuses_Count	Created_At
15846407	TheEllenShow	Ellen DeGeneres	California	http://www.ellentube.com	Comedian, talk show host and ice road trucker. My tweets are real, and they're spectacular. http://www.ellentv.com	46922050	36720	103498	253	10907	2008-08-14 03:50:42
23375688	selenagomez	Selena Gomez	Los Angeles	http://smarturl.it/GoodForYouSG	Philippians 4:13 oh and remember you are awesome.	32267984	1269	133897	44	3906	2009-03-09 00:16:45
19397785	Oprah	Oprah Winfrey		http://www.oprah.com	None	28783937	242	89429	135	10509	2009-01-23 15:18:34
50393960	BillGates	Bill Gates	Seattle, WA	http://www.gatesnotes.com	Sharing things I'm learning through my foundation work and other interests...	24315040	168	115782	4	1766	2009-06-24 18:44:10

Overall Daily Tweet Graph for 'Ebola'



Analysis of User Vs Disease Related Tweets