Reimagining Our World at Planetary Scale: The Big Data Future of Our Libraries

Kalev Leetaru
Georgetown University

Kalev Leetaru, "Reimagining Our World at Planetary Scale: The Big Data Future of Our Libraries" (2016). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316501

Reimagining Our World at Planetary Scale: The Big Data Future of Our Libraries

Kalev Leetaru, Senior Fellow, Center for Cyber & Homeland Security, Georgetown University

This is a transcript of a live presentation at the 2016 Charleston Conference.

Kalev Leetaru: Thank you so much. It is a true honor to be here with you all today. Most of my career over the last 20 years has focused on how do we reimagine society through the massive amounts of data that we have with us? And I want to open with this project here.

This is a project I did several years ago. I approached the Internet Archive and said, "You know, I want to reimagine the book." What if we thought of books not as containers of text but as the world's greatest art gallery? Lots of people in the past have explored pulling images off of digitized books, but I wanted to see what would happen if we did that at scale. So, in the end, we took 600 million pages of books dating back 500 years pulled from over 1,000 libraries worldwide, and took the existing OCR that the Internet Archive had already done, and pulled off every image off of each page and all the text that surrounds that. What we can do is really fascinating things, finding fascinating examples like this, which came from, I forgot the era, but it was apparently a wedding (referencing image on slide). We can find fascinating things like to remember that 100 years ago, mail by train, taking your letter and sticking it in a train going by, was this incredible way of speeding delivery of messages compared obviously to 100 years prior. Being able to walk through 500 years' worth of imagery, and the image in the top left there is a woodcut from 500 years ago, but the image to the right is a very famous one, the telephone and communication, and it's really fascinating. One of my favorite books, *1921*, there was an illustration of a teenager using a telephone, and the book said, "In the future, we predict that teenagers will actually make use of the telephone heavily."

It's very funny to really see it visually, to be able to see how things are depicted, and famous illustrations through time, but also images like this. So, these two images, these are from Emblem Books, and the two books I think are about 100 or maybe 200 years apart; one from a library here in the U.S. and one from a library in Europe. Now, historically a historian would likely never have encountered both of these two images from these two books from these two libraries. It's unlikely that someone would have stumbled upon those two, but in the context of having all these digitized, machines can go through and identify similar images that we can find these two images and say, "You know what? The later artist probably copied this earlier one. This is too similar to be coincidence." So, the power that data offers us to really understand imagery. And this image here, starting in the back left, this is one image selected from each year, so this is 500 years of images. The top left is the year 1500, one image selected. The next year is 1501, 1502, 1503, and on down, so you're seeing in one image 500 years of how imagery in books has evolved over time. If you zoom into this, it is actually quite fascinating how the styles have changed, to see the introduction of color, to see all these fascinating things like the scientific revolution go through there. Again, there's really powerful ways of visualizing and understanding society.

But then I want to come to this image. So, what you're seeing if you look in the lower left, you can see the year, and what this shows is, I applied tools of algorithms to go through all the books published in a given year and pull out every location on Earth mentioned in those books in a given year. What you're seeing here, and so we're at 1870, 1880, 1890, watch and in a moment, you'll notice that the map is getting bigger and bigger and bigger, and then it just pops. It suddenly shrinks. Well, that's 1923. That's copyright resuming. And so this map is a very powerful map because it really demonstrates to us how much copyright is impacting digitization and especially data mining, and the conclusions that we can draw from all this data because we know so much between 1800 and 1922, all of this incredible amount of data is available to us, but come 1923 when so much of the digitization efforts have basically ended at that point, we basically loose—we understand more of what happened 200 years ago than we do in the last 70 years. So, this is just something to keep in mind as we think about the world.

But, I also want to bring that same theme to this map here. So, the map you're seeing at the top there, it's basically a grid of the world looking over the last three years at any location on Earth where

https://doi.org/10.5703/1288284316501

there was a large number of "Tweets" sent from during that period. This is an animation by month over the last three years from 2012 to the end of 2014. You'll notice Twitter expands over time, but then it kind of halts. It kind of freezes in time in the middle of 2013. The bottom two graphs show you basically a scaled 1% sample showing the number of "tweets" per month and the number of unique users sending "tweets" per month, and this is very important because we think often times today in our big data world, we grab for data. We say, "Wow, we're drowning in data!" Hundreds of, actually in this case, billions and billions and billions of "tweets." Look at all this data that shows a society, but to never lose track of the fact that yes, we have these incredible volumes of data, but in the case of something like Twitter, this is only reflecting a really tiny portion of the world, and that view of the world has really frozen in time. For folks like me that are data miners, this is really important because social media is not frozen. Social media is just exploding across the world. I think Facebook has now like 1.7 or 1.8 billion users now or something like that, but that data isn't accessible to us. The majority of Facebook posts are private. The majority of social media as a whole is set to private. So Twitter was kind of the great public experiment. We have all this data to try and peer into society's soul but to never lose track of the fact that that never really let us look beyond a very small portion of the world, and that view is really frozen. I think this is something that we all too often lose track of it is, who are we really hearing from? What are the views that we're really seeing across the world?

And this was an example, just looking at this red box there. Some of you may recall last year there was this big battle between Yemen and Saudi Arabia. In Yemen, they fired a missile upon Saudi Arabia. Well, that box there represents Yemen, and all the white dots here are anywhere that any "tweets" were sent in that period. You see that up north of that in Saudi Arabia there are a lot of people "tweeting" there and in Yemen not so much, so we're really missing half the story. Even as the stories we can tell through data become richer and richer, the comprehensiveness of that is shrinking, and I think that is a critical thing to think about. You think about 20, maybe just 10 years ago, you might be looking at say a gigabyte of data out of a 10-gigabyte data set. Today we're still looking at a gigabyte, but it is out of hundreds of terabytes of data. The tools that we use, the questions that we ask haven't really scaled beyond what we were asking a decade ago. It always amazes me that I think it was last year was the 50th anniversary of the "Origins of Dialogue," and it always strikes me that half a century later the majority of the ways that we interact with information is still through keyword searches. Half a century later, we're still searching through keywords, and that has always struck me as that is not the natural way to search information and to think about the world.

Even when it comes to things like news, so this is an example. This is March of 2015, and this is looking at—orange are all the locations mentioned in BBC news coverage. Green are all the locations mentioned in *New York Times* coverage, and for those of you that work closely with the social sciences, you'll know that there's a whole area of social sciences, political science, communications, and so on where they hire a large team of humans to look through newspapers and write about what they are seeing there and try to catalog the world. But what we find, of course, is typically that involved looking at something like *The New York Times*, one single paper, and really data is very powerful because it allows us to really visualize and demonstrate that the world is this incredibly rich place, and no single source gives us a perfect view of society.

And that is why—so what I really want to talk to you today is a project that I call the GDELT Project, which is an open data project supported by Alphabet Jigsaw, formerly Google Ideas, and a number of others. And this project is really about how do we catalog human society? How do we take all of that data that is out there today and try to understand what is happening around the world, specifically in a particular area, so that we can try to bring the world closer together? Because if you think about all the things, if you open up say *The New York Times* today or a paper here in Charleston, you're only seeing the smallest microcosm of society. How can we reach across the world and try to understand what is happening in the entire planet so that we can tell stories that aren't being told, that aren't being heard?

Like any good project, this begins with data. So, news media, over 100 worldwide news media, print, broadcasts, and web, over 100 languages are processed; 65 of those are live translated. And in partnership with the Internet Archive, we sent them

a list every night of all the online media we find around the world, and they archive them into the way-back machine to preserve online journalism. This is one of the largest initiatives in the world to preserve online non-Western journalism because so much, I mean, take for example the coup in Turkey. Shortly thereafter, almost half of the major media in Turkey disappeared overnight, so being able to really preserve that for perpetuity. Television—there's another collaboration. We're looking at that. Academic literature—this I think also really demonstrates the ability of data miners to collaborate with publishers. So, this is a project—among others, JSTOR provided access to a lot of its holdings to be able to data mine and to be able to look through academic literature, to be able to understand what have social scientists and humanists and ethnographers and linguists and so on, what have they really written about the world over the last half century? What can we understand about that? How was the output of academia compared to say the output of the news media? You know, how are these really combining? Books, of course, and then imagery, which I will come back to in a second.

Of course, if you look across the world, most of the world's media and most of the world's information is in a language other than English. What you're seeing here is a "dot" at every location which GDELT monitored information from or about over about a six-month period of last year. It is color-coded by the primary language. Gray is English, and you'll notice there's not a lot of gray on this map. So if you're only looking at English-language material, which is what most data mining does and what most American scholars tend to focus on, you're missing most of the world here. Yes, there are plenty of news sources and other materials published in France that are published in English, but those don't reflect local views or local events or local things that are happening there. So, just always remember how important it is to look at information. When you're trying to understand the world, you have to turn to all the world's languages, and in this case using machine translation. Now machine translation is far from perfect, but it's good enough. For example, I don't speak a word of Thai. So, if I see a Thai language article, I can't tell you if this is say a local cricket match score or if this is an anti-government protest. Being able to leverage the technologies that we have today to really try to understand the world, you can imagine bringing in all of this information from across the planet, and anything that is not in English translating that in English, what could you do with all that? Why do you want to do that? What is useful about all this? Well, two things.

The GDELT Project looks at two things: It takes—one, it tries to understand the physical events. So, what is this newspaper article telling me? Is this reporting on a riot, a protest, a coup, a peace appeal, a diplomatic exchange? So, that is the event data set. So much of what we are interested in and around the world is not physical events. It is narratives. It is emotions. It is beliefs. Those that work in the humanities world know how important that is there. It is not enough to say, "Hey, a coup took place in Turkey." We all know that. What we care about is how are people reacting to that around the world? What are the emotions and the narrative beliefs?

What can we do with this? Well, for a physical event data set, if you take a large fraction of the world's news media and you process that, and you make a list of every protest reported around the world over the last day 40 years from the data set, you can take countries and actually plot the intensity of protest activity and ask questions, like when Greece was having its issues, what was that, a year ago? I saw this article, I think it was on CNN, that said, "Greece Undergoes Worst Violence in its Modern History" and to being able to look back and say, "Yeah, there was certainly some unrest but nothing compared to five years ago, and certainly nothing compared to times before that." So, really being able to contextualize things because, again, we have such short memories and being able to really understand the world around us or being able to do automated alerts. To be able to say not many people were watching Burundi in December of last year, but to be able to say hey, all of a sudden something was happening there that we need to pay attention to. So, being able to put that in front of say peace builders to be able to—again people on the ground, obviously people in Burundi know what's happening there, but being able to tell that story to the world. Because when something happens in Turkey, the world listens. If something happens elsewhere in the world, it's not making the front page of say *The New York Times*, so how do we tell these stories? How do we tell what's happening?

And I won't go into this slide, but you can do things like actually look at the cycles of world history. We can actually build mathematical models over

hundreds of millions of events that have taken place across the world and actually find similarities there; cycles, patterns that repeat themselves that allow us to start really peering into the soul of society. We can do things like map. Now this is very old data. That's why it is very sparse. This is I think from 3 years ago, but being able to actually map the world, actually watch the world go by moment by moment, to say give me a map of all the protests happening right here in the world. What are the grievances? What are the things that are driving society? Is the world becoming more democratic or more autocratic around the world? Being able to ask these questions, to say you know what? We've had things like news media and other data sets. These are not new things. We've had them for very long time, but for the first time, we had the data to be able to ask questions of that. To be able to map things beyond, things like emotions, but to go beyond positive and negative. This is actually a map of anxiety, and this is the big trend mine. The surge you're seeing there is the U.S. government shut down. This becomes very powerful because this was the case where there wasn't a lot of positive or negativity because you had a lot of people say, "Hey, great! I get a month vacation! Fantastic!" You had other people say, "This is horrible! Our government is shutting down! This is awful!" So, in terms of positive/negative, both sides, there weren't—it was sort of like the election today, who is going to win the election? Well, depending on which candidate you're going for, it's a good or bad thing. But anxiety is very high because either side you're worried about things so being able to really ask deeper questions about society and being able to look at things like Ebola coverage.

So, this is a map, a timeline of coverage of Ebola in American television news, and you can see that thousands of people are dying across the continent, and there is no news coverage. Television is never mentioning it. It is not until the two Americans get it that suddenly it is worth American media covering it. But that bottom graph is fascinating: That is how positive or negative the coverage is. You'll notice that after the Americans get it, media coverage becomes actually strangely positive about Ebola. That is because the first coverage is very uplifting. CNN's first article about Ebola was very uplifting. It said if Ebola makes our shores, we are all going to die. It's going to wipe us all out instantly. Very reassuring coverage. But once the Americans get it, it's like, "Don't worry; now that Americans have it, it is American medicine to the rescue! This will all be over soon and everyone will be cured." So, being able to really capture that, and do those that study journalism know this happens? Being able to quantify this and actually be able to put that in a map and, of course, with the election right now doing a lot of work and being able to actually count how many times each candidate is mentioned on television each day across the United States and by market, you know, across geographically different across the country and what people are saying about the candidates. So, making—being able to say how are they contextualizing? Is Donald Trump the savior of our nation, and or is he the devil incarnate? Is Hillary a criminal, or she this incredible figure that is going to transform our nation? There is no right answer to these things. This all depends on who you are and where you are in this country, so being able to contextualize that, to peer inside all of that today, this is really fascinating to be able to see just how polarized the nation has become.

Being able to map things, the map in the upper right is very interesting. So, that is a map of wildlife crimes, so things like illegal fishing or poaching or so on, and what was fascinating about that is when I first went to make this map, I was told by a lot of folks in the community, "Don't bother trying to map what the media talks about with wildlife crime. Nobody talks about that. You're not going to find anything." Because people are saying, well, *The New York Times* doesn't cover it unless there's some huge ivory seizure. I don't really see much in *The New York Times* each day. Well, that's true, but local media in local communities do cover this every day, so always remember that even if we don't see things in the information that we consume, then just remember that globally there is so much information that is out there, and the dark map, the network diagram in the center, that was something that BBVA, the big bank in Spain, they wanted to see how was Russian media portraying the sanctions, the economic sanctions against Russia. So being able to say, "Well, how is the media within a country portraying something compared to what we're seeing out here?"

And in particular, I want to show this map. This was something. I wanted to map global happiness. How happy or sad are people of the world? This is a question people have asked for eternity, but I said, well, what if we take, in this case 200 million articles published last year in over 100 different languages, 1.3 billion mentions of locations on Earth, and about three quarters of a trillion emotional assessments off

of that. What I did was say, well, every article worldwide that mentions Paris, what is the average happiness or sadness of all of those mentioned across the entire planet and do that city by city across the world? You end up with a fascinating map like this. There are so many fascinating stories here. You also start reflecting media. Certain country's media reflects more of a negative tone toward things, but in particular in Europe, it's a little bit hard to see on the screen, but you see these really strong negative tendrils go through Europe. That's a domestic reaction to the refugee crisis that is occurring there. So, being able to see how domestic countries are reacting to world events and just seeing that on a global stage is very powerful. By the way, I should mention that was one line of codes, so that was a tool that Google makes called Big Query. It's a super database platform. One single line of SQL query, 60 seconds. One minute later and you've processed that much data. So, really we've moved to the ability where if we have a question, we say I wonder or what if? The fact that we can actually answer those questions or be able to peer into that media and say who are all the people talking about a particular issue? Like, CCS is a clean energy technology and being able to look inside of that and say what is this environment look like? What does this landscape look like? And this is something I want to conclude on.

So, this is a project—Google, like many companies, is doing deep learning, neural networks, artificial intelligence. They are doing some amazing work on having AI systems. They can look at images and catalog what are the objects in that image? What are the activities in that image? Does this image depict violence? What are the logos, the locations? Can we look at the background of that image? And look across the planet and say can we estimate the location of this image? And that is very powerful because we can use this to do incredible things. What you see in the upper left, I asked it for all the images of trash around the world. This is actually something I will be putting out shortly. This looks across the world in real time and tries to estimate basically environmental conditions, so litter conditions, air pollution, and so on across the planet but in particular looking at things like flooding. One of the things that we do is when a natural disaster like Hurricane Matthew passes through, being able to take all the images emerging from those areas and say, we don't want just presidents at podiums. We just want the images that show the destruction so we can send those off to first responders so they know what's happening across these areas. So again, being able to move beyond text to really understanding the imagery of the world. And the fact that we have the tools today to look through, and so far, we've processed 175 million images since this past January from every country on Earth in real time. The ability to use tools to really try to understand the narratives of the world, and I want to conclude on this image. This image really, by the way, this is a real image; this is not Photoshopped. This is the NOA science ionosphere. It is really cool. It is a 6-foot acrylic sphere hung from the ceiling with projectors around it. This is one of my data sets on it. This really kind of summarizes what I think is so powerful about today, that we have so much data. We have the tools—we've always had more data than we can deal with, but we have the tools for the first time, the machines, the algorithms, etc., to make sense of that data, to let us peer into the soul of society and really understand it in ways that we've never been able to do before. Thank you so much.