Lightning rounds: A look inside the world's largest initiative to understand and archive the world's news

Kalev Leetaru, senior fellow, The George Washington University and founder of the GDELT Project |
Oct. 14, 2016 | Charles E. Young Research Library, UCLA |
Dodging the Memory Hole 2016: Saving online news

KALEV LEETARU: ***[00:07]*** Well thank you so much for having me here today.

[00:12] So, this is my vision. This is the GDELT Project. The idea of the GDELT Project is, how do we take the world's information and try to catalog what's happening around the world, moment by moment, and how are people reacting to that? The events, the emotions, the reactions; both the textual and the visual narratives that surround that. How do we preserve that, as well? So, like any good project it begins with data, and it has many different data sets that feed into it. So, news material. This is print, broadcast and web, over 100 languages supported, covering every country of the world. Sixty-five of those are live-machine translated, in real time, as they arrive.

[00:44] In collaboration with the Internet Archive, as Mark helpfully put in there, to preserve as much of that online material, and we're constantly ramping up the volume of the material they're preserving for us. And the goal of that is to really say, especially with online material — again, so much of this material is ephemeral, especially when you move outside the US, as we know obviously in Turkey, most recently. Television — so again, we have television partnerships throughout the world, but also within the United States. Collaboration again with the Internet Archive to process about 100 stations here.

[01:08] Academic literature. So, this is a collaboration: JSTOR, DTIC Core sites, Internet Archive. Processing 21 billion words covering 70 years to really provide the context to why things happen around the world. Books. If you want to go backwards in history, human rights material and imagery, which I'll come to in a second. So, obviously, as we all know in most of the world English is not their primary language. This is a map of where GDELT has pulled information from the first six months of last year, color-coded by the primary language. Gray is English.

[01:39] So obviously, a lot of the work that has been done historically in preserving online media has focused on English. And, as you can see from the grayish fonts there, if that's all you're preserving, you're missing most of the world here. Now, of course, broadcast and print are very crucial because in many of the areas of the world, print and broadcast are the only source, especially broadcast, are your only sources into those areas. But again, this really shows how important it is to reach beyond, and especially to do things like translation or process things in languages other than English.

[02:07] In terms of preserving online news, about 140,000 articles that we monitor each day disappear within about two months. And this is a very important thing for us because it's not evenly distributed across the world. For example, in Nigeria, under Goodluck, articles that would criticize

the government's response to Boko Haram tend to have a very, very short shelf life indeed. Obviously, Turkey; many countries around the world. But also even here within the United States, a news outlet goes bankrupt, a small local outlet goes up. Also, again, news outlets unfortunately are more and more pulling material. Also, as we've seen, many major publishers now treat content as very fluid. They have no fear of editing material in real time. And I also wanted to point out in Nepal here: so, a lot of efforts that have been done to date have really focused on things like manual archiving, saying, "Hey, a major disaster just took place, let's start preserving now."

[02:59] You miss that. So, in Nepal, for example, we've grabbed about 667,000 articles over the year following that earthquake, really capturing especially the local perspective on things. We can do a lot with that material, so we can do everything from mapping human rights violations to looking at refugee flows, MANPADS and all kinds of other things. Also, using the Google Cloud Vision API, we're cataloguing a large fraction of all that imagery worldwide each day. About 150 million images to date from almost every country in the world. We're extracting out locations, activities, objects, facial emotions, OCR in 80 languages. And also things like violence and so on.

[03:33] So the ability to look not just at the textual narrative, but really for the first time be able to peer into that visual narrative.

[03:39] And finally, I want to leave you with this map. Once you have all this data, what can we do with the world's news? How do we think about it in new ways? This is a very simple map to look at global happiness through the eyes of the media. Two hundred million articles in 65 languages, 750 billion emotional assessments, including things like anxiety, fear of the future, optimism and so on; 1.5 billion mentions of location and 150 million images.

[04:00] And being able come up with a map of what does that show us, at least through the eyes of the world's news media. What are we seeing? Again, being able to make maps like this and say, well why do we see more red here? What does that tell us? Whether this is legitimate for society or whether the media in that particular area focuses and portrays things in a particular light? Thank you very much. Come to my talk this afternoon to hear more on this.