MIT CENTER FOR CIVIC MEDIA

LOGIN

CATHERINE D'IGNAZIO

Research Affiliate

Catherine D'Ignazio is the person behind that really cute baby. She is an Assistant Professor of Data Visualization and Civic Media at Emerson College who investigates how data visualization, technology and new forms of storytelling can be used for civic engagement.

Professor D'Ignazio has conducted research on geographic bias in the news media, developed custom software to geolocate news articles and designed an application, "Terra Incognita", to promote global news discovery. She is working on sensor journalism around water quality with PublicLab, data literacy projects and various community-educational partnerships with her journalism students. Notably, she co-organized a hackathon at the MIT Media Lab called "The Make the Breast Pump Not Suck!" Hackathon.

Her art and design projects have won awards from the Tanne Foundation, Turbulence.org, the LEF Foundation, and Dream It, Code It, Win It. In 2009, she was a finalist for the Foster Prize at the ICA Boston. Her work has been exhibited at the Eyebeam Center for Art & Technology, Museo d'Antiochia of Medellin, and the Venice Biennial.

Professor D'Ignazio is a Fellow at the Emerson Engagement Lab and a Research Affiliate at (and alumna of) the MIT Center for Civic Media.

kanarinka.com

BIG DATA, NEWS AND GEOGRAPHY: RESEARCH UPDATE

Submitted by kanarinka on October 3, 2013 - 12:01pm

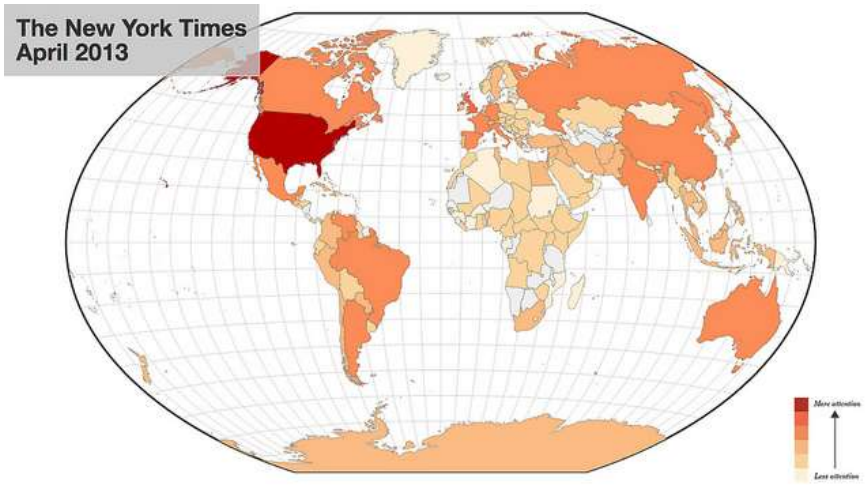
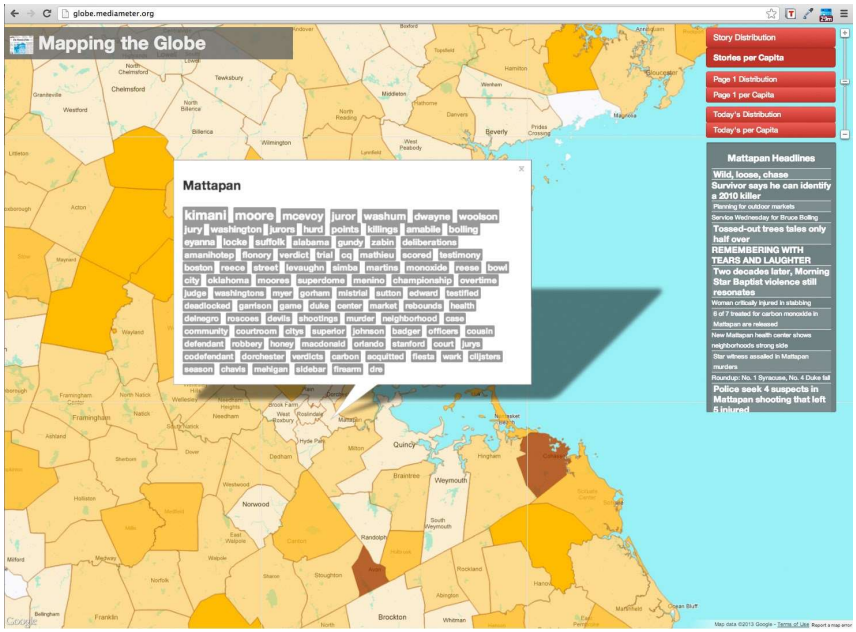


Image: One month of the world according to the New York Times

At the Center for Civic Media we do a lot of demos for the Media Lab in which we synthesize our work for visitors, sponsors and classes. Recently in our demos we talk about the three big three questions that our work in quantitative media analysis addresses: *The What, Where and Who of Attention*. We are interested in *what* topics are being discussed by mainstream media sources as well as social and participatory media. We are interested in *where*, geographically speaking, gets more (or less) attention and how media influences space and place. And we are interested in *who* gets to speak in the complicated new media ecosystem.

For *What* we have work like Attention Plotter by Erhardt Graeff (part of the Controversy Mapper project) that seeks to map how a local news story like the Trayvon Martin case becomes a full-blown media firestorm. Nathan Matias has been working on *Who* by analyzing the gender breakdown of internet media and by creating personal media interventions like Follow Bias. And as for *Where* - well, that's what this blog post is about.



Mapping the Globe

We began mapping the news in the Fall of 2012 with a project called Mapping the Globe. Our friends at the Boston Globe's GlobeLab gave us access to their alpha API with around two years of news stories. It is updated every day with new stories and all stories are geolocated by the reporter that submitted them (that's high quality geodata!). We mapped those stories across the Greater Boston and MA region to get a sense of both the quantity of coverage (Which neighborhoods/towns get more coverage than others? Why?) and more qualitative measures of coverage (When the Globe talks about Mattapan, what does it talk about?). You can explore the data yourself here and read about our preliminary results from that project.

Mapping the News in a "Big Data" Way

How can we scale up the maps and analysis from Mapping the Globe? What kinds of research questions could we ask if we could map twenty five years of articles from the New York Times? Or map a decade of coverage from all US political blogs? What kinds of patterns in geographic coverage might we see by comparing how different sources "see" the world? What regions are overrepresented in English language news? Where are the blank spots on the map? To be able to scale up like this we needed two things:

1. Lots of news data
2. Automated geolocating of news stories

We've got #1 covered by our project Media Cloud which archives 27,000+ RSS newsfeeds on a daily basis. #2 was a trickier problem which the rest of this blog post will address.

Geoparsing - Algorithmically mapping large news data sets

What made Mapping the Globe feasible was its high-quality, manually submitted geodata. To scale this up we needed an automated solution that would be able to ingest large amounts of news articles and reliably locate the country a particular story is about. This is called "geoparsing" - taking unstructured text, extracting the places from it (entity extraction) and figuring out which places in the world they correspond to (geographic disambiguation). There is an excellent paper by Kalev H. Leetaru that explains this process in detail.

There are a number of services and products out there that do geoparsing. Last Spring we compiled a list of all of the geoparsing tools we could find. Ideally we needed a tool with the following characteristics:

1. Accurate: We would need to assess the tools on their accuracy in comparison with manually entered geodata.
2. Free (as in beer and as in speech): Since we need to geolocate news at scale and for research purposes we wouldn't be able to afford a service like Yahoo's PlaceSpotter which runs $8.00/1000 queries. We are also committed to building Media Cloud in a way that others can reproduce our work so we needed something that wouldn't be hidden away in a corporate blackbox.
3. Open source and modifiable: We wanted to be able to integrate geoparsing into our toolchain for Media Cloud articles and to be able to tune it as necessary to more accurately geolocate text from news articles.
4. Runs locally: We would prefer a standalone technology to an API that runs over the network to save time and computing power.

From this wish list, we chose three technologies to evaluate:

1. Yahoo PlaceSpotter: Along with MetaCarta's Geotagger, this is known as an industry standard and enterprise solution in this space. Although it is too expensive to be a viable option for us, we decided to evaluate PlaceSpotter to see what kind of results we could reasonably expect from geoparsing news articles.
2. OpenCalais: OpenCalais takes unstructured text and turns it into structured content: named entities, places, products, companies. It is free (or mostly free) and runs as a web service.
3. CLAVIN: Named after Cliff Clavin of Cheers fame, CLAVIN is a geoparser written in Java based on Stanford's well regarded NER parser (for entity extraction) and the geonames.org gazetteer of over 8 million world placenames. CLAVIN is free and open source.

Place mentions vs "Aboutness": What are we testing?

These technologies all have limitations when compared with a human. They pick up place mentions, disambiguate them to one place on the globe, and give us a latitude and longitude, but they don't actually tell us what place an article or text is "about". It is the latter that we were actually interested in so we would have to come up with a way to measure both the technologies' performance *and* our ability to get to "aboutness" through a list of place mentions.

We took a naive approach to aboutness and decided to evaluate whether *frequency of mention* would correspond to aboutness at the country level. Our hypothesis was that if an article mentioned one particular country most often (or places in that country most often) then the article was most likely about that country.

Here are the things we measured in our evaluation:

1. **Precision:** What percent of the of the place references the technology found were accurate place references (versus false positives)? If an article mentioned "Georgia" 7 times and the tool found 6 references but 2 of them were actually for a person named "Georgia" then the precision would be 4/6 and there would be 2 false positives.
2. **Recall:** What percent of the actual place references did the technology detect? In the previous Georgia example, the recall would be 4/7. You can also think of this as "completeness".
3. **Geographic disambiguation:** What percentage of the time did the technology disambiguate the place mention to the correct place in the world? This is also referred to as the "Springfield Problem". Was the tool able to correctly guess that Springfield meant Springfield, MA, and not Springfield, VT?
4. **Aboutness:** What percentage of the time was the most frequently mentioned country the place that the article was about?
5. **F1 measure**: The F1 measure is a weighted average of precision and recall in information retrieval systems and an overall measure of a system's accuracy. The best possible score is 1 and the worst is 0.

Our Test Data Set

We wanted to use a high quality, hand-coded data set of news articles from different sources in order to evaluate these technologies. Different sources were important because different news organizations have different style guides, article lengths and reading levels. There might be variation in the accuracy of a geoparsing technology based on the media source. We chose to use a set of 75 articles—25 from the New York Times, 25 from the Huffington Post and 25 from the BBC. They were randomly sampled from the month of February 2013.

Humans as the Gold Standard

We needed to test the technologies performance against the "right" answers as determined by a human. However, even humans don't necessarily agree on what places are in unstructured text. Though place would seem to be a simple problem, in fact it turns out to be quite difficult. For example, take this sentence from one of the articles we coded,

> *"All of the police officers on the Shrine were Tajiks or Uzbeks from northern Afghanistan who said they enlisted and came to the Pashtun south because they believed in their country and its government; they were nationalists."*

How many places are there here? "The Shrine" is geographic data even though we don't have enough context from this sentence to geolocate it. The demonyms "Tajiks" and "Uzbeks" are giving us geographic data even though they refer to people. Is "northern Afghanistan" the correct place reference or just "Afghanistan"? And "Pashtun south" refers to the southern part of where the ethnic Afghans live which does not necessarily correspond to national borders.

These questions are hard enough that we knew that humans would not always agree on them so we needed a measure for how often humans agreed with each other on 1) place mentions in news articles and 2) overall "aboutness" at the country level for each article.

We came up with some rules for hand-coding the place edge cases:

1. Demonyms ("French", "Tajik") count as places
2. Organizations ("Columbia University") do not count as places
3. Geographic features without context to locate them do not count as places ("The Shrine")

4. Regions count as places ("The Middle East")
5. Rivers and geographic features with context count as places ("The Helmand River")
6. Military sites count as places ("Camp Leatherneck", "Guantanamo Bay")
7. Sports teams, organizations and universities with places in their names count as places ("University of Virginia")
8. Street mentions and addresses count as places ("24th Street")

HUMAN AGREEMENT AVERAGES	
Average human agreement on place mention	92.23%
Average human agreement on aboutness of article	96.15%

Luisa Beck and I separately hand-coded the set of 75 articles using these rules. Once we compared our results, it turns out we agreed 92% of the time on what constituted a place in unstructured text. We agreed 96% on "aboutness" at the country level i.e. what country an article was actually about. This gave us a good baseline for the peak possible performance of any given geoparser.

The Results

GEOPARSING AVERAGES	RECALL: Accuracy at picking up place mentions from unstructured text	PRECISION: % relevant place mentions	Accuracy at Disambiguating places	ABOUTNESS: Frequency of place mention corresponds to "real country"	F1 measure
Yahoo Placespotter	69.50%	87.70%	96.27%	69.02%	0.78
OpenCalais	53.03%	96.78%	90.28%	59.57%	0.69
CLAVIN	63.78%	94.25%	89.91%	75.30%	0.76

We ran our tests with the 75 articles from three different news sources and averaged the results which you can see in the table above. You can also look at our full results breakdown by media source. It's important to note that we filtered out places that humans did not agree on and only ranked the technologies against places that two humans had separately agreed on. Here are some of the relevant outcomes:

- Yahoo consistently outperforms OpenCalais and CLAVIN on recall and geographic disambiguation. Overall it picks up more place references than the other two technologies.
- Yahoo has a higher rate of false positives - incorrectly identifying text as a place—whereas CLAVIN and OpenCalais have high precision scores.
- Because of that, the F1 scores for CLAVIN and Yahoo are comparable (0.76 and 0.78 respectively) whereas OpenCalais scores lower (0.69)
- All of the technologies did relatively well with geographic disambiguation (solving "The Springfield Problem") with Yahoo leading them at 96%. This means that Yahoo gets the right Springfield 96% of the time.
- Using frequency of mention as a measure of "aboutness" of an article actually worked pretty well for CLAVIN. We were able to correctly locate an article at the country level about 75% of the time using CLAVIN. Not bad! And then we made it even better by improving on CLAVIN's disambiguation strategy (see below).

Fine-tuning CLAVIN's disambiguation strategy

While writing a server application for CLAVIN, Rahul Bhargava, Research Specialist at the Center for Civic Media, noticed that he could improve on CLAVIN's disambiguation strategy. Using his new class to solve the "Springfield problem", we have now bumped our aboutness results up to 85% instead of the 75% we previously saw.

The way our disambiguation strategy works is by taking multiple passes through the article and making contextual guesses about what places are being referred to and by privileging countries (since that's the administrative level that we are most interested in at the moment). For example, the first couple passes pick up large areas ("the Middle East", "Africa") and country names. The next couple passes look for places that have a population and are located in countries that found in a previous pass. And the last couple passes try to make educated guesses about the place if we are still uncertain.

CLAVIN wins

With our improvements to CLAVIN's disambiguation strategy our implementation of CLAVIN can accurately locate the country of a news article 85% of the time. This is just 9% less than human agreement on the "aboutness" of a news story.

In what we have been affectionately terming our "Great Geoparsing Bake-off" CLAVIN emerged as the clear winner. Its performance is almost as good as Yahoo's more expensive enterprise service and it met all our other criteria. Because it is open source we can integrate it into our toolchain and tweak it to perform better with news articles.

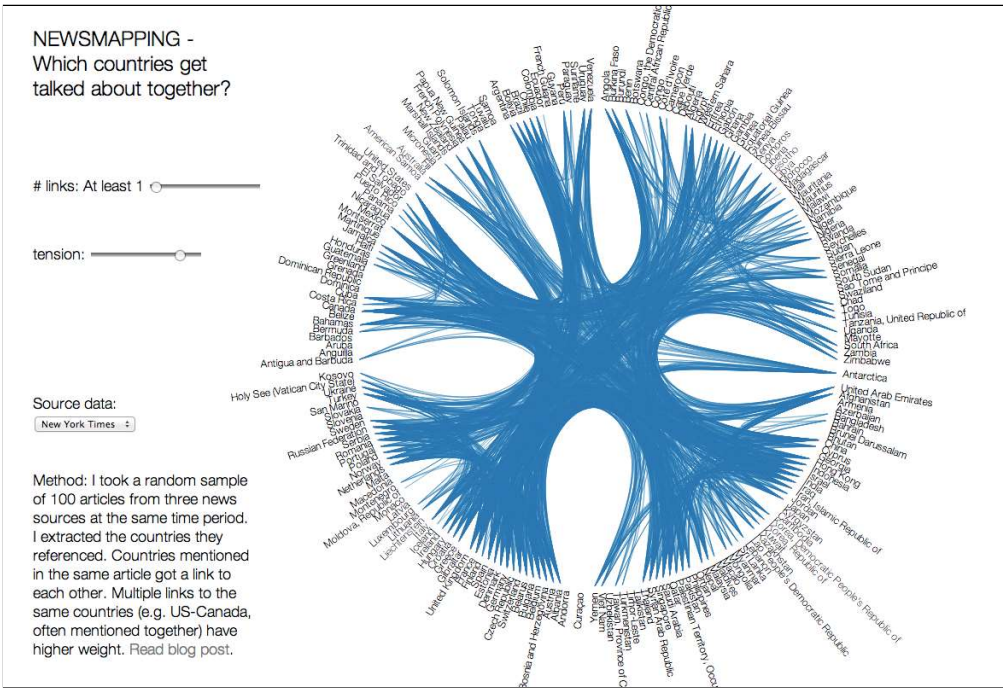
Further improvements and questions

We have ideas for how we could get even better results from our geoparsing with CLAVIN. Here are some things we will investigate:

- Add demonym disambiguation to CLAVIN ("French" and "Chilean" should resolve to France and Chile)
- Make sure CLAVIN's disambiguation favors large populated cities versus higher administrative levels (e.g. "New York" should resolve to "New York City" over New York state)
- Work on places with two or more words in their name (e.g. Washington D.C.) which currently trip it up
- Refine our naive "aboutness" algorithm. Frequency of mention turns out to be a fairly good approximation of aboutness at the country level. But we could also take into account the confidence of each mention, the location of the place mention in the text (weight places mentioned at the beginning of a text higher), "roll up" many place mentions at a lower administrative level like states into a higher administrative level like country, and so on.
- Support other languages. It would be pretty fancy to be able to cross language barriers and create comparative maps of the world according to Chinese media versus US media for example. In practice this simply means downloading a different version of the Stanford NER parser (or training one if it doesn't exist for that language).

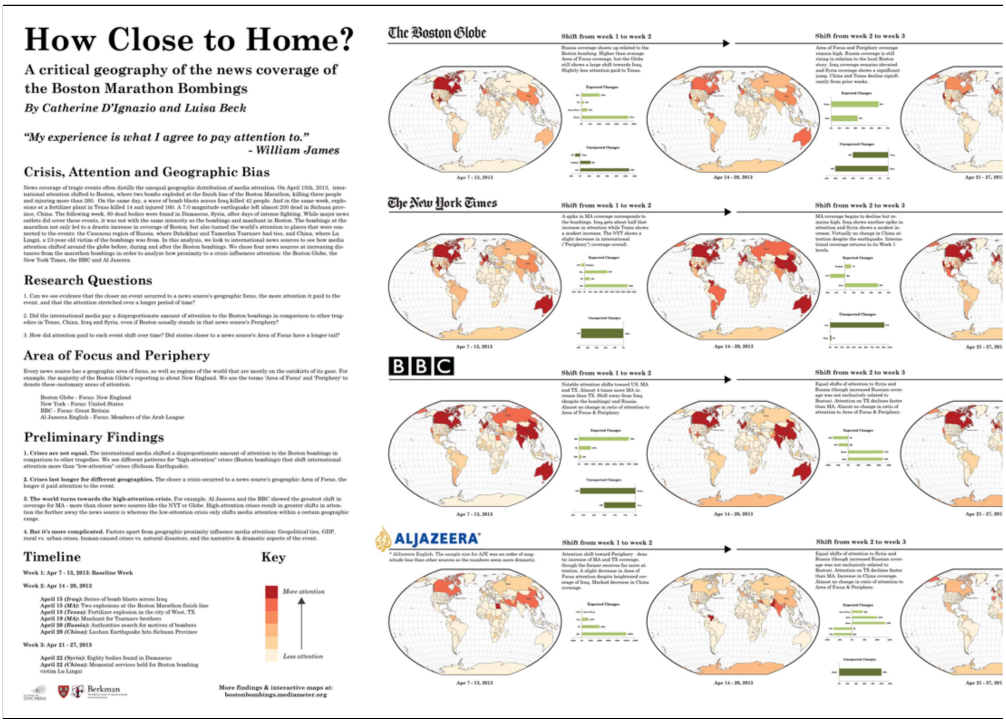
What we can do now

Now that we have solid evidence that we can reliably geolocate articles at scale 85% of the time, the door is open for some exciting critical geography work. We've done a couple of small experiments in this realm. For example:



Mapping International Relations Through the News: What countries get talked about together? I took a small sample of 100 articles from three news sources and grabbed the places they referenced. If two countries were referenced in the same story then they got a link. Multiple links to the same countries (US-Canada) have higher weight. You can play with the # of links slider to see countries that most frequently get referenced together.

Comparative News Maps: We compared a single month of news coverage (April 2013) between seven news sources, including mainstream media, digital native media and blogs. To make comparisons about long-term patterns of coverage, we would obviously need more than one month of data, but this points in the direction of the kinds of maps we could make in the future along with running regressions to see if factors such as GDP and population play into longitudinal patterns of international news coverage. Ethan Zuckerman did an analysis similar to this in his paper Global Attention Profiles in 2003 that we are looking to update and refine.



Topic-based News Mapping: Luisa Beck and I analyzed international news coverage for four major news sources (Al Jazeera English, BBC, NYT, and the Boston Globe) during the weeks before, during and after the Boston Marathon bombings. We were interested in why the marathon bombings got so much more attention in relation to other crises and tragedies that happened that same week in Syria, China, Texas and Iraq. You can explore our maps online and read our analysis.

Our next major step with this will enable users to search and map all 27,000+ Media Cloud sources according to keywords. For example, where in the world are people talking about "guns"? About "peace"? About "government shutdown"?

Make your own maps

To make your own news maps you just need to install the CLAVIN version tagged 0.3.3, install our CLAVIN-Server branch 0.3.x and start feeding it news articles. Currently, you need some programming knowledge in Java to be able to make your own news maps. In the coming months we will be connecting CLAVIN-Server to Media Cloud and building an interface for running your own topic-based queries against the many millions of articles in Media Cloud. At that point making your own news maps will only require you to come up with a good research question and plug in the right search queries and news sources. Get in touch with us if you have some good ideas for news maps.

Related Previous Posts

- Mapping the Globe: Initial Research Into Regional Media Attention in Massachusetts
- Mind the Map: Toward a Handbook for Journalists
- How Close to Home?: Crisis, Attention and Geographic Bias
  - Blog post
  - Interactive crisis maps
- News Mapping Post: A Comparative Experiment in Mapping the News