### **Visual cross-platform analysis: digital methods to research social media images**

**Supplemental Information: Data Collection**

The data driving our article’s intervention is derived from an analysis of climate change related content across five different platforms. While an exhaustive discussion of the particularities of the data is outside of the article’s scope, this appended section will briefly outline our methods of data collection as well as some of the justifications behind them. We first discuss the initial platform selection and then detail the different data-gathering methods that were employed for each of chosen platforms.

*Platform Selection*

Platform selection was based on a combination of pragmatism and platform popularity. Pragmatic choices related to issues of data access and availability within the summer school’s dedicated time span of one week. Platform popularity also weighed into the selection of the individual platforms to ensure comprehensive and representative coverage (high-membership mainstream platforms) and better comparability (platforms of scale). On the individual platforms, different entry points and methods were chosen for data collection. These were developed based on the specific medium’s affordances, usage norms, and narrative cultures, although some pragmatic choices were necessarily made along the way.

*Twitter*

To collect data from Twitter, the Twitter Capture and Analysis Toolset was used. The tool has been collecting climate change related tweets since 2012 (including tweets that mention climate, drought, flood, global warming, and globalwarming) and had reached a total dataset of 134.402.719 tweets at the time of the study (June 2017). Due to the very large dataset, studying all-time data was not possible within the given timespan. Therefore, a two 2-week period was chosen: May 24 2017 until 7 June 2017 (time frame covering US President Trump’s announcement of the US exit from the Paris Agreement). Furthermore, tweets were narrowed down to those explicitly mentioning “climate change” or climatechange (as either text or hashtag). The final dataset contained 418.111 tweets from 186.238 individual users.

*Facebook*

We approached data collection on Facebook through the analysis of 10 popular (high-membership) community pages related to climate change: Campaign Against Climate Change, Climate Central, Climate Change Guide, Climate Outreach, Climate Reality, Greenpeace International, NASA Climate Change, Met Office UK, Friends of the Earth International and Nature Climate Change. These were selected as significant and authoritative activist and scientific sources of knowledge. The decision to center our data around these community pages reflects the fact that Facebook’s social networks are connected at friendship and group membership nodes as opposed to hashtags. We used these pages as an entry point to Facebook because the API does not permit keyword searches across the entire platform (in contrast to the other platforms in this study).

Having chosen 10 pages, we scraped each of them for all-time images using the Netvizz Facebook API tool. The tool provided us with a tabular file including the metadata for each image including: image id, image name, time of creation, link to the Facebook page, link to the image URL, like count, reaction count and comment count. The 10 individual data sets were ranked by highest reaction count to determine the top 10 most popular/engaged with images across the 10 pages. Using the [DownThemAll](https://wiki.digitalmethods.net/bin/edit/Dmi/DownThemAll?topicparent=Dmi.MakingClimateVisibleFacebook) Firefox Module, the top 10 images from each page were downloaded. Two pages, Met Office UK and Friends of the Earth International, were then excluded from the data set on the basis that their image content did not directly relate to climate change.

*Reddit*

Given that Reddit content is organized by sub-reddits – forums dedicated to specific themes or topics – we drew on sub-reddits to collect images on the platform. Sub-reddits were identified via 2 separate search terms: “climate change” and “global warming.” Using the Google search engine, we searched for the following query: “global warming” OR “climate change” site:reddit.com. Drawing on the first two pages of Google results, we then identified and selected six relevant sub-reddits: r/science/ filtered by “flair:environment” or “flair:env”, r/GlobalWarming, r/environment, r/climate, r/climatechange, and r/climate\_science. To scrape the images on the sub-reddits, we utilized the scraping tool Big Query (Google Cloud Platform). The output of this tool is a csv file with the following meta-data: title, score, url, domain, permalink, thumbnail link, num\_comments, created\_utc, author, subreddit.

We collected two different datasets from our sub-reddit corpus with Big Query, both of which spanned one year, from May 2016 to April 2017. Our first dataset narrowed the selection of six sub-reddits down to three that prominently contained posts with associated images/thumbnails, and represent Reddit communities of various sizes: r/science/ filtered by “flair:environment” or “flair:env” (17,263,410 subscribers), r/environment (243,087) subscribers, and r/GlobalWarming (1,699 subscribers). This dataset generated 16897 hits. The second dataset relied on searching for relevant posts across sub-reddits as opposed to collecting all of a particular sub-reddits posts. We used Google Big Query to find all posts the had the words “climate change” or “global warming” in either the title or body text (selftext) across all of the six originally chosen subreddits, generating 18962 hits. We then used Google Vision API to code the 5000 most engaged with images in each dataset. Lastly, with Table2Net and Gephi, we generated a network map that shows the relations to the tags that Google Vision API applied to the images.

*Instagram*

Content on Instagram is organized around individual user accounts and feeds as well hashtags that group the posts of many users together. We focused on hashtags in order to collect a broad and representative sample, and because this approach was facilitated by Visual Tagnet Explorer. This tool queries a single hashtag and outputs a gephi file of cotag networks as well as a csv file of the following metadata: username, location, filter, likes, comments, caption, image thumbnail, and link to the original post. We wanted to compare Instagram posts before and after the announcement on June 1st 2017 of the US withdrawal from the Paris Agreement. We initially queried #climatechange on June 26th, but the liveliness of the hashtag and the limitations of the tool and platform API resulted in a corpus that only reliably stretched back to June 10th. Therefore we decided to query #parisagreement, a popular hashtag found within the #climatechange dataset. This did provide a corpus that straddled the withdrawal announcement. The corpus totalled 17,320 media items which dated from February 3 – June 26, 2017.

We then used Gephi’s modularity algorithm to sort the cotag network into 16 clusters. We identified the most mentioned tag of each cluster to explore in more depth. These tags were: nature, environment, America, art, savetheplanet, vegan, trump, repost, love, politics, world, heatwave, project wildlife, Philippines, climatecamp, and waterislife. We drew on the data from the csv file that was connected to each of these tags in order to collect the thumbnails that were affiliated with them. Using [DownThemAll](https://wiki.digitalmethods.net/bin/edit/Dmi/DownThemAll?topicparent=Dmi.MakingClimateVisibleFacebook) we then downloaded all of the images from each cluster. We also sorted the images within each cluster based on the engagement numbers (likes and comments summed) of the posts. Then, we conducted a visual content analysis by plotting the top ten most engaged with images from each clusters and from the whole the dataset. To do that, we relied on the ImageSorter tool.

*Tumblr*

DMI’s open access [TumblrTool](http://labs.polsys.net/tools/tumblr/), which utilizes the Tumblr API to retrieve posts tagged with a specific term, was used to collect data from Tumblr. For this study, we queried the term “climate change” over the past year by running individual queries for each month. The overall range of the dataset was 27/06/2016-26/06/2017 (the year leading up to the study). We retrieved two types of results: tabular files containing basic descriptions of the retrieved posts (including user name, URL of the image, etc.) and co-tag files. The search strings returned various types of posts, featuring text, quotes, links, videos and/or images. We filtered the results to keep only posts with images. We ranked the 12 individual datasets (1 for each month) to identify the 10 posts that received the most engagement on the platform over each monthly period. Engagement hereby refers to both “likes” and “re-blogs”. In a subsequent step, we identified the top 10 posts of the past year. Using the DownThemAll Firefox Module, we downloaded these images for further processing.