Alexa Source Coding Guidelines

We are coding for two variables - source country and source type. Below are guidelines for how we should do this coding. We are primarily interested in the top news websites, for the purposes of geoparsing news articles to locate what countries they are about.   If the url includes a path, treat it as if it doesn't (ie. http://facebook.com/blah/blah should be treated as if it said http://facebook.com).

Source Country
Visit the website and identify what language it is in - this may narrow it down to one country immediately. Failing that, look for an "About" or "Contact" page to identify what country they operate from. Fill in the 2-letter ISO3166 code for the country.

Source Type
We want to categorize each source by what type of content they are. We have a number of options:

- **aggregator**: Text news sites that gather content or abstracts from other sources and show them all together. These types of sources do not produce much original content at all. (ie. news.yahoo.com)
- **broadcast**: Text news sites that are primarily broadcast media outlets that serves a large number of people. (ie. cnn.com)
- **magazine**: Text news sites that primarily operate a print magazine but also have an online presence. (ie. forbes.com)
- **mismatch:** A website that is in the wrong category, or isn't a news site. For instance Alexa categorizes things in a taxonomy, but we might not agree. (ie. shutterstock.com in the News category, or youtube.com because it isn't primarily a news site and it isn't text-driven)
- **newspaper**: Text news sites that are primarily print newspaper companies. The website is the online presence for the newspaper. (ie. usatoday.com)
- **online**: Text news sites that were launched and live primarily online - "digital native" publications. (ie. huffingtonpost.com)
- **weather**: Sites that are primarily weather websites. (ie. wunderground.com)
- **wire:** Text news sites that are wire services that sell their stories to other media sources. (ie. ap.com)