I went to a “scrapeathon”. This is what happened

March 19, 2015

Credit: @knuutila

Why scrapeathons?

Yesterday evening I went to a community event at Kings Cross Impact Hub hosted by OpenCorporates. I first heard about OpenCorporates when the founder, Chris Taggart, was speaking at Open Knowledge Conference 2013. Their main idea is that – as we live in a world of ubiquitous but more or less opaque organizations – it’s best for all of us when we know more about these organizations and how they are connected. For this purpose they started building up a database of global corporate data.

Their idea is to take all the data that is already out there – in tables on websites, in PDFs, in databases – to extract it, and to bring it into a consistent format ready for everyone to be tapped. As this isn’t a task that could be done by some programmers in their spare time, they run community events where they invite people to write programs that are able to extract data from a specific data source. These programs, which convert human-readable into machine-readable data, are called scrapers. We therefore might call these events scrapeathons.

Scraper 101 (for non-scrapers)

As I do not code, I was more than happy that the OpenCorporate folks did not expel me right away, but quite the contrary, tried to explain to me how the process works. That’s what I remember: Most of the data online could be easily copy&pasted into some kind of database. That might sound useful in the first place but would urge us to repeat the process every time the data is updated. Therefore people write programs which are custom-fitted to a specific data source and which can then be applied to the data source on an iterative basis. Once the scraper is up and running it is transferred to OpenCorporates (using a tool called Turbot) for them to use the scraper to fill and update their databases.

Apparently you need quite different scrapers for different data sources. In the introduction to the event Peter Inglesby of OpenCorporates showed us how to extract data from a table on a website. You basically have to identify the table in which the data sits (using the web developer tool of your browser) and then write a short e.g. Python command to extract all lines of the table and convert them to a JSON string (the data format they prefer at OpenCorporates). To me this looked quite feasable and I could imagine that I try something like this on an upcoming event myself. What I heard from the other people, data sources like PDFs on the web or databases might be more tricky to crack open. At some point in time I left the coders code and moved over to the beer and nibbles corner of the room. However I guess that most of the people managed to finish one scraper during the 2 to 3 hours they were there.

What's in for me?

As I already figured out in my first case study in Berlin the practice of scraping has a quite central meaning to the development of open data institutions in a city. Especially in the very early days of open data, when there was hardly any awareness of the concept within public bodies, a lot of data was already out there on the web sitting in some kind of tables of just plain text. This data was neither completely closed, nor was it officially open. To exploit and resolve this ambiguity activists started to just scrape the data and republish it as open data on their homemade data portals. As a result many public bodies then decided to open up their data themselves – to retain at least a bit of (quality) control and not to "lose their face" again. Tonight I will take part in a meetup of the geospatial-startup-community to figure out the role of scraping and crowdsourcing in their work to redefine the distribution of information on our day-to-day environment. Stay tuned.