

Návrh vizualizace - Katastrofy a nehody 1998 - 2008

VIKBB47 | Vizualizace informací

374201 | Lukáš Strouhal

397476 | Jan Hrabal

Jaký problém má vizualizace osvětlit?

Naše datová vizualizace zobrazuje na mapě světa nejrůznější přírodní katastrofy a jiné větší nehody nepřírodní povahy. Cílem je ve srozumitelné podobě předložit uživateli dataset, který obsahuje údaje o jednotlivých událostech, a tyto údaje znázornit s pomocí přidané geografické hodnoty. Je tak možné vysledovat, která místa na světě jsou postižena častěji či méně často určitým typem neštěstí a zároveň se dozvědět, jaké události byly v minulosti svým rozsahem výjimečné oproti jiným.

Kde je zdroj dat?

Data pochází z otevřené mezinárodní databáze katastrof [EMDAT](#), kterou tvoří Centre for Research on the Epidemiology of Disasters (CRED) pod hlavičkou World Health Organization (WHO). Zprostředkovaně jsme dataset našli pomocí služby [Infochimps](#).

Jaké jsou v datech problémy?

Základní dataset jsme získali jako TSV soubor, se kterým jsme dál pracovali převážně v MS Excel a Google Refine. Původně obsahuje přes 18000 záznamů - událostí od roku 1900 do roku 2008. Ke každé události jsou uvedeny informace o časovém rozmezí trvání, zemi a místu, typu a upřesnění události (případně i se jménem) a také o počtu zabitých, katastrofou ovlivněných a také o škodách. Množství řádků nemělo kompletní popisy, chyběly například údaje o počtu zabitých nebo bližší geografický popis. Rozhodli jsme se některé údaje ignorovat a dále s nimi nepracovat, konkrétně počet ovlivněných a vyčíslení škod. Tyto sloupce byly v porovnání s ostatními nejčastěji prázdné.

Vzhledem k velikosti tabulky jsme se taky rozhodli pro další práci s omezeným datasetem a to pro údaje, které označují události let 1998 ž 2008 - přes 8000 řádků.

Prvním problémem byla úprava časových údajů. Ty byly ve formátu celého čísla (např. 02041999 nebo 052004). Za pomoci funkcí v Excelu se nám podařilo z tohoto nepříjemného formátu, který často nenesl kompletní informaci o přesném datu, dostat rozumnější vyjádření.

Hlavním problémem bylo získání zeměpisné polohy (délky a šířky), která je nutná pro správné umístění události na mapu. Zprvu jsme využili online [geocoder](#), využívající služeb Yahoo!, který vracel souřadnice zadaného místa v textové podobě. Ty jsme následně sloučili s datasetem. Metoda se však později ukázala jako znatelně chybová a umístění událostí na mapě nekorespondovalo se slovním popisem z původního zdroje. Využili jsme tedy vlastního kódovacího mechanismu [CartoDB](#), ovšem za cenu, že ne všechna data nám úspěšně projdou až na mapu. Kromě těch, které z různých důvodů nebylo možné

rozpoznat (chybějící údaje, úplně špatně zapsaná poloha, neurčitá označení jako sever, jih a podobně.) nám CartoDB menší část dat ignorovalo úplně - důvod se nepodařilo zjistit. Z původního datasetu zhruba 8000 dat je tedy na konci na mapě vizualizovaných zhruba 6.500-7000 událostí.

Ještě před samotným kódováním v mapovém softwaru jsme se pokusili pro zvýšení úspěšnosti opravit údaje o poloze pomocí Google Refine - především překlepy, zkratky, jinak zapsané názvy těch stejných měst, slova jako between a near nebo ujasnění polohy v případě, že je uvedeno více měst.

Proč ne holá data?

V případě takto rozsáhlého datasetu je práce pouze s vlastní tabulkou velmi chaotická a nepřehledná. V momentě, kdy by chtěl uživatel najít informace o konkrétní události v konkrétním místě nebo oblasti, může využít filtrování, ovšem získaná data neuvidí v souvislosti s daty ostatními. Stejně tak geografická poloha vyjádřená slovem není vhodná pro snadné a rychlé pochopení reálné polohy.

Návrh vizualizace

Náš návrh vychází z možnosti vytvářet vlastní interaktivní mapy za pomoci služby CartoDB. V zásadě zobrazuje nakódovaná místa událostí pomocí kružnic, které jsou schopny vyjádřit dva údaje - barvu (typ události) a poloměr (škálovaně počet mrtvých). Tohoto vzhledu jsme dosáhli úpravou CSS pro naši mapu a pokusili se tak rozšířit běžnou klikací funkcionalitu. Kružnice jsou průhledné, aby zmírnily problém, kdy je více událostí na jenom místě a plné body by se mohli úplně překrýt. Také (a nejen) proto je mapa zoomovatelná. Po rozkliknutí události se objeví další údaje o konkrétním bodu.

Při realizaci nás napadlo několik dodatečných možností, jak návrh vylepšit, na které naše schopnosti bohužel nestačily. Jednou z nich je možnost do mapy implementovat časovou osu, která by umožnila zvolit si období, pro které chci události vidět, nebo si pustit postupné zobrazování tak, jak bude na ose datum narůstat. Také by nebylo špatné, kdyby mapa obsahovala přepínače zobrazení podle typu události, případně podle dalších charakteristik (třeba počet mrtvých od do a podobně). Samozřejmě je možné do již existující mapy přidat zpočátku vynechané údaje od roku 1900. Nepřidali jsme je z důvodů neznámého omezení nakódovaných dat od CartoDB (viz výše) a z důvodů přehlednosti - bez možnosti filtrace či časové osy by mapa byla nejspíše příliš plná.

Příklad realizace: <http://cdb.io/1c7oFpE>