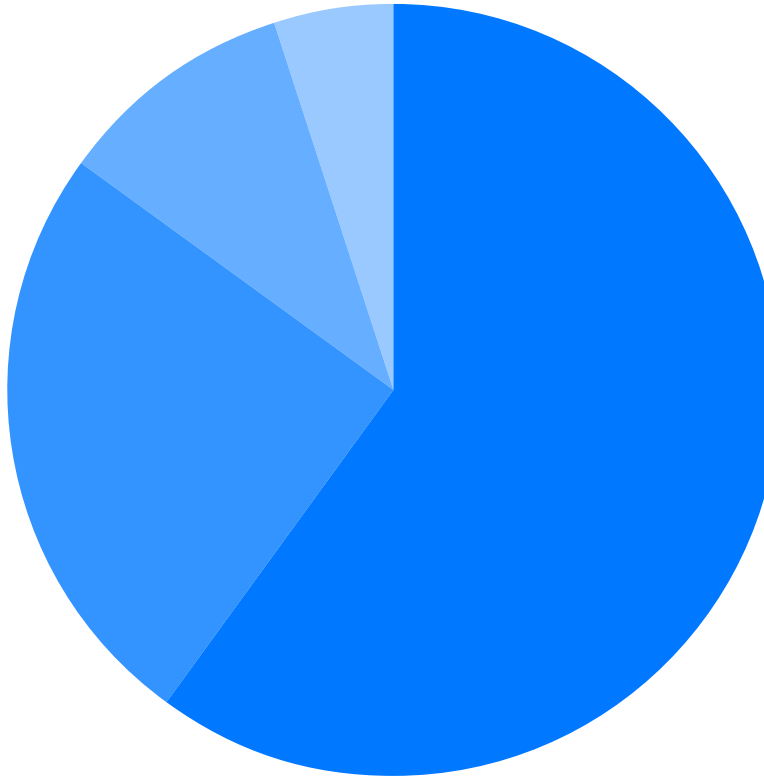
Nadácia otvorenej spoločnosti

príručka dátovej žurnalistiky

potenciálni čitatelia:

60 % novinári
25 % študenti žurnalistiky
10 % watchdogové organizácie
5 % nadšenci, programátori



obsah

- predslov 05
- úvod 07

1. proč je datová žurnalistika důležitá 09
2. datová gramotnost ve třech krátkých krocích 17
3. tipy pro práci s čísly 23
4. ako získavať informácie 27
5. extrakce dat z webu 35
6. tipy pro práci s internetovými zdroji 41
7. vizualizace jako tažný kůň datové žurnalistiky 45
8. tipy pro vizualizaci dat 53
9. dátová žurnalistika na slovensku 57
10. mladý datažurnalistický talent, ondrej proksa 61

- doporučené nástroje 66
- doporučené zdroje 67
- autoři the data journalism handbook 68

Vydala Nadácia otvorenej spoločnosti – Open Society Foundation v rámci projektu Frontline Clubu Slovensko „Dáta menia žurnalistiku“.

NADÁCIA OTVORENEJ SPOLOČNOSTI

Poslaním Frontline Clubu Slovensko je predovšetkým podpora kvalitnej a objektívnej žurnalistiky. Frontline Club preto podporuje vzdelávanie študentov v oblasti dátovej žurnalistiky a tiež prepájanie študentov s profesionálnymi médiami. Nadácia otvorenej spoločnosti vyhlasuje každoročne súťaž Novinárska cena, v rámci ktorej oceňuje kvalitnú slovenskú žurnalistiku.

Projekt Dáta menia žurnalistiku je podporený spoločnosťou GOOGLE.

Toto vydanie príručky vzniklo v rámci vzdelávacieho projektu Dáta menia žurnalistiku v spolupráci so spoločnosťou GOOGLE.

Google™

predslov

Keď sme v decembri 2013 spúšťali prvý ročník Ceny Google pre mladé talenty dátovej žurnalistiky, rozhodli sme sa ho zamerať na študentov rôznych univerzít a mladých ľudí do 30 rokov. Vylúčiť zo súťaže profesionálnych novinárov a médiá sme sa rozhodli z dvoch dôvodov. Pokúšame sa vzbudiť záujem a hľadať talenty na univerzitách, kam táto téma rozhodne patrí. Do mediálnych redakcií sa dostávame cez porotcov tejto kategórie, zástupcov médií, ktorí majú k dátovej žurnalistike blízko. Pevne veríme v inšpiráciu, ktorú študenti svojimi nápadmi môžu redakciám priniesť či už vo forme nejakého nápadu, alebo širšieho prístupu k téme.

Toto upravené vydanie Príručky dátovej žurnalistiky je vytvorené s ohľadom na skúsenosti a žurnalistickú prax na Slovensku. Smeruje rovno na univerzity, kde tento rok v rámci projektu Dáta menia žurnalistiku spúšťame pilotné semináre dátovej žurnalistiky, ale aj k ďalším záujemcom o dátovú žurnalistiku.

Spolu s viacerými novinármi, ktorí si dátovú žurnalistiku vyskúšali a čiastočne sa jej stále venujú, citlivo vnímame rôzne bariéry, ktoré stoja jej rozvoju v ceste. Aj napriek nim, máme spoločný záujem o jej podporu. V princípe sa všetci zhodneme, že systematická dátová žurnalistika dnes v slovenských médiách neexistuje. Zásadné aktivity a dátové projekty niektorých mimovládnych organizácií však vytvárajú skvelé podmienky na využitie dát aj v žurnalistike. Príručka, ktorú práve čítate, smeruje preto aj k novinárom a profesionálom v slovenských médiách a ponúka im inšpiráciu a konkrétne postupy pri práci s dátami.

Možno sa všetci čoskoro stretneme v jednom datasete, tak nech má dobre vyčistené dáta, ukáže nám nové súvislosti a zaujímavé príbehy.

Ľubica Stanek
Nadácia otvorenej spoločnosti – Open Society Foundation
Bratislava, december 2015

úvod

S pokrokem fotografických technologií se v druhé polovině 19. století objevil nový žánr novinářské fotografie; žurnalistika získala kvalitativně zcela nový nástroj a všichni společně jsme získali nový pohled na svět kolem sebe.

Podobná situace se opakuje dnes, řekněme od přelomu tisíciletí. Díky technickému pokroku je stále větší část našeho světa popsaná pomocí čísel, od světové ekonomiky a řízení státu až po vztahy mezi lidmi na sociálních sítích. Není tedy divu, že se před pár lety objevila nová disciplína jménem data journalism, datová žurnalistika, která nám nabízí nový pohled na svět; tentokrát nikoliv hledáčkem fotoaparátu, ale displejem počítače, prostřednictvím čísel.

Podobně jako dobrá fotografie není jen otrockým obrazem okamžiku, ale nositelem příběhu, podstatným postřehem o našem světě, i kvalitně zpracovaná data jsou mnohem víc než suchá statistika určená k založení do zaprášených šanonů. Data bez nadsázky hýbou světem a zachraňují životy.

Čísla si neprávem vysloužila pověst něčeho neprostupného a zároveň nudného. Datová žurnalistika dokazuje, že série grafů a vizualizací může mít sílu fotografií pořízených na válečném poli při nepoměrně větším nasazení. Zároveň jsou čísla součástí našeho každodenního života a jistá „datová gramotnost“ bude brzy nezbytnou výbavou nejen specializovaných, ale i běžných novinářů a jejich čtenářů. Těm všem je určen následující text.

1.

proč je datová žurnalistika důležitá

Zeptali jsme se několika předních praktiků a zastánců datové novinařiny, v čem podle nich spočívá význam oboru.
Tady jsou jejich odpovědi:

Informační filtr

Dokud bylo informací poskrovnu, většina novinářského úsilí spočívala v jejich shánění. Dnes, když je informací nadbytek, je důležitější jejich zpracování. To má dvě hlavní části: analýzu, ve které se snažíme v nekončícím přívalu informací zorientovat a najít smysl, a prezentaci, kde čtenářům překládáme to důležité a relevantní. Datová žurnalistika se podobá vědě: veřejně popisuje své metody a dává k dispozici dostatek podkladů, aby se výsledky daly ověřit.

Philip Meyer, emeritní profesor
University of North Carolina at Chapel Hill

Budoucnost žurnalistiky

Datová žurnalistika je budoucnost novinařiny, novináři se musí umět vyznat v datech. Dřív jste sbírali informace po barech (ostatně to nejspíš občas funguje dodneška), ale do budoucna budete muset analyzovat data, mít ty správné nástroje a umět vybrat to důležité. Udržet informace ve správné perspektivě. Ukazovat lidem, jak věci zapadají dohromady, co se ve vaší zemi děje.

Tim Berners-Lee
zakladatel World Wide Webu

Nový nástroj

Datová žurnalistika nabízí prostředky, které tradiční žurnalistika postrádá: nástroje pro hledání v digitálních zdrojích, analýzu dat a jejich vizualizaci. Nechce tradiční novinařinu nahradit, ale doplnit, rozšířit.

Dnes, kdy se většina zdrojů digitalizuje, mají novináři možnost a povinnost být těmto zdrojům blíž. Internet otevřel možnosti, o kterých se nám ani nezdálo. Datová žurnalistika je začátkem evoluce našeho současného systému práce pro online svět.

Datová žurnalistika hraje v každé redakci dvě důležité úlohy: jednak je zdrojem nových námětů, které z jiných zdrojů nedostanete, a jednak umožňuje novinám plnit jejich úlohu hlídacího psa. Zvlášť v obdobích finanční nejistoty jsou oba tyto cíle pro noviny zásadní.

Z pohledu regionálních novin je datová žurnalistika naprosto nepostradatelná. V redakci máme takový postřeh, že „volná dlaždice před vaším domem je důležitější než zahraniční státní převrat“. Ta dlaždice se nedá minout, má bezprostřední vliv na váš život. Podobně mají regionální noviny bezprostřední vliv na své okolí, takže vzhledem ke všudypřítomné digitalizaci musí jejich novináři umět hledat, analyzovat a vizualizovat data.

Jerry Vermanen
NU.nl

Odpověď na datové PR

Měřicí nástroje jsou dnes snadno dostupné a jejich cena klesá; společnost se na všech úrovních zaměřuje na efektivitu a výkon. Tyto dva faktory se navzájem posilují a vedou k tomu, že se při rozhodování stále víc hledí na kvantitativní ukazatele, trendy a možnosti.

Firmy přichází s novými a novými metrikami, které je ukazují v lepším světle. Politici zbožňují řeči o poklesu nezaměstnanosti a růstu HDP. A kauzy Enron, Worldcom, Madoff nebo Solyndra jsou důkazem novinářské neschopnosti prohlédnout závoj čísel. Konkrétní čísla mají ve srovnání s jinými fakty větší šanci na nekritické přijetí, protože se kolem nich vznáší jakási aura důstojnosti, přestože jsou třeba kompletně vymyšlená.

Zběhlost v práci s daty vrací novinářům schopnost kriticky reagovat na čísla. Doufejme, že jim vrátí také některá území ztracená ve válce s PR odděleními.

Nicolas Kayser-Bril
Journalism++

Nezávislá interpretace oficiálních dat

Japonsko je země, která v digitální žurnalistice doposud zaostávala, což se bolestně projevilo zejména v roce 2011 po drtivém zemětřesení a následné katastrofě v jaderných elektrárnách prefektury Fukušima.

S hrůzou jsme zjišťovali, že vláda ani odborníci nemají žádná důvěryhodná data o způsobených škodách. Když vláda před veřejností zatajila data ze systému SPEEDI, týkající se rozptylu radioaktivních látek, nebyli jsme je připraveni zpracovat ani v případě, že by je někdo vynesl. Dobrovolníci začali sbírat radioaktivní data pomocí vlastních přístrojů, ale bez znalosti statistiky, interpolace a vizualizace. Novináři musí mít přístup ke zdrojovým datům a musí se naučit nespoléhat na jejich oficiální výklad.

Isao Matsunami
Tokyo Shimbun

Data jsou náš život

Kvalitní datová žurnalistika je dřina, protože kvalitní žurnalistika je dřina. Musíte umět sehnat data, pochopit je a najít v nich kvalitní námět. Občas narazíte na slepou uličku, občas žádné jiné ani nejsou. Ostatně – kdyby stačilo jen zmáčknout to správné tlačítko, nebyla by to žurnalistika. Právě proto naše práce dává smysl. A ve světě, kde data tvoří stále větší a větší část našich životů, je datová žurnalistika nepostradatelná pro svobodnou a spravedlnou společnost.

Chris Taggart
OpenCorporates

Časová úspora

Novináři nemají čas na to, aby data přepisovali ručně nebo se je snažili vytahat z PDF souborů. Když se naučíte trochu programovat nebo víte, kde sehnat pomoc, je to velké plus.

Jeden reportér z novin Folha de São Paulo pracoval na článku o městském rozpočtu a volal mi, aby mi poděkoval za městské účty, které jsem dával na web. Pro mě to byly dva dny práce, zatímco on už je prý kvůli článku ručně přepisoval tři měsíce. Podobné to bylo s organizací Contas Abertas, která monitoruje dění v parlamentu: řešení jejich „problému s PDF“ mi zabralo 15 minut a 15 řádek kódu, zatímco pro ně představovalo měsíce práce.

Pedro Markun
Transparência Hacker

Nezbytná součást výbavy

Podle mě je důležité zdůraznit, že datová žurnalistika je především žurnalistika. Analýza a vizualizace dat nemají smysl jako samoúčelné cvičení, ale pouze jako nástroj, který nás přiblíží k pravdě o dění v našem světě. Schopnost analyzovat a interpretovat data vidím jako nezbytnou součást dnešní novinářské výbavy, nikoliv jako samostatnou disciplínu. Ve výsledku jde vždy především o kvalitní novinařinu, schopnost vyprávět příběh tím nejvhodnějším způsobem.

Datová žurnalistika je další možnost, jak objevovat svět a hlídat představitele moci. Vzhledem k rostoucímu množství dat je dnes důležitější než kdykoliv předtím, aby novináři zvládali i datovou žurnalistiku; ať už sami, nebo ve spolupráci s někým druhým.

Hlavní sílu datové žurnalistiky vidím ve schopnosti získat informace, které by se jinak hledaly nebo dokazovaly jen těžko. Dobrým příkladem je článek, ve kterém Steve Doig analyzuje škody způsobené hurikánem Andrew. Steve propojil informace z databází stavebních úřadů s informacemi o škodách způsobených hurikánem a zjistil, že část škod byla způsobena uvolněnými stavebními předpisy. V roce 1993 za tento článek dostal Pulitzerovu cenu; je velkou inspirací a důkazem toho, co všechno je možné.

V ideálním případě můžete s pomocí dat najít anomálie, body zájmu, něco překvapivého. Tady data fungují jako stopa, indicie. A přestože jsou data zajímavá sama o sobě, psát jen o nich nestačí. Úkolem vás jako novináře je také vysvětlit, co znamenají.

Cynthia O’Murchu
Financial Times

Adaptace na změny v našem informačním prostředí

S novými digitálními technologiemi se ve společnosti objevují nové zdroje informací a nové metody jejich šíření. Datová žurnalistika se dá chápat jako snaha médií o adaptaci a reakci na změny v našem informačním prostředí. Do tohoto rámce zapadá i nový, interaktivnější způsob vyprávění příběhů ve více rozměrech a vrstvách, díky kterému mohou čtenáři prozkoumat data, na kterých je článek postaven, a zapojit se do procesu jeho vzniku a kritického hodnocení.

César Viana
University of Goiás

Datový vesmír

Z naší digitální stopy se dá rekonstruovat celý náš život. Co čteme, kam a kdy cestujeme, co posloucháme, naše první lásky, první kroky našich dětí, dokonce i naše poslední přání – to všechno se dá sledovat, digitalizovat, ukládat a analyzovat. Z tohoto datového vesmíru si můžeme odnést příběhy, odpovědi a myšlenky, které bychom z osobních svědectví při nejlepší vůli neposkládali.

Sarah Slobin
Wall Street Journal

Otevřená data pro zpětnou kontrolu

Na web často dáváme kromě vizualizací také data ke stažení. Čtenáři tak mají možnost data prozkoumat pomocí interaktivní vizualizace nebo si je stáhnout a zpracovat podle potřeby sami. Jaký to pro nás má význam? Zvyšuje to průhlednost Seattle Times. Předkládáme čtenářům stejná data, ze kterých odvozujeme naše závěry, často zásadní. A kdo té možnosti využívá? Rozhodně naši kritici, a kromě nich všichni, kdo se o příslušný článek hodně zajímají. Publikovaná data fungují i jako zpětná vazba – kritici i běžní čtenáři nás mohou upozornit na něco, co jsme přehlédli, co by šlo dál vytěžit. Pokud chce člověk dělat novinařinu, na které záleží, tohle všechno jsou plusy.

Cheryl Phillips
Seattle Times

2.

datová gramotnost ve třech krátkých krocích

Stejně jako slovo gramotnost označuje schopnost získat a kriticky posoudit psané informace a vyjadřovat se srozumitelně v psaném projevu, spojení datová gramotnost označuje schopnost získávat znalosti, kriticky uvažovat a srozumitelně se vyjadřovat prostřednictvím dat. Patří sem nejen jistý pojem o statistice, ale také schopnost práce s velkými objemy dat a představa o tom, jak vznikají, jak je navzájem propojit a jak je interpretovat.

Floridská nezisková škola žurnalistiky Poynter Institute nabízí v rámci svého projektu News University předmět Matematika pro novináře, ve kterém se studenti učí bezpečně pracovat například s procenty nebo aritmetickým průměrem. Zajímavé je, že tytéž koncepty se v těsném sousedství učí také žáci pátých ročníků základních škol, tedy děti ve věku 10–11 let.

Pokud novináři potřebují pomoc s matematikou základní školy, musí mít průměrný newsroom k datové gramotnosti daleko. Což nutně vede k problémům – jak může novinář zpracovat čísla o změně klimatu, když neví, co je interval spolehlivosti? Jak může napsat článek o příjmech domácností, když si plete aritmetický průměr s mediánem?

Zároveň ale novinář k práci s daty nepotřebuje titul z matematiky. I pár jednoduchých postřehů může z čísel udělat lepší článek. Jak říká Gerd Gigerenzer, profesor na Ústavu Maxe Plancka, lepší nástroje samy o sobě nedělají lepší žurnalistiku, pokud nejsou podepřené vlastní úvahou.

I bez větších znalostí matematiky nebo statistiky můžete udělat krok k lepší datové žurnalistice – stačí si položit následující tři jednoduché otázky.

Odkud se data vzala?

Fantastický růst HDP

Když chcete někoho omráčit, nejlépe se to dělá daty, která jste si sami vymysleli. Možná je to evidentní, ale kašírovat se dá i tak diskutovaný údaj, jakým je například HDP. Někdejší britský velvyslanec Craig Murray ve své knize Murder in Samarkand popisuje údaje o HDP Uzbekistánu, které vznikají na základě intenzivního vyjednávání místní vlády s mezinárodními organizacemi. Jinými slovy: nemají nic společného s místní ekonomikou.

Vlády si HDP jakožto hlavní ukazatel výkonu ekonomiky hlídají kvůli dani z přidané hodnoty, která pro ně představuje hlavní zdroj příjmů. Když vláda žije z jiných zdrojů než DPH, nebo když nezveřejňuje svůj rozpočet, nemá důvod sbírat podklady pro výpočet HDP a je pro ni jednodušší výsledné číslo prostě vymyslet.

Věčně rostoucí křivka zločinu

„Zločinnost ve Španělsku vzrostla o tři procenta," píše El País. „Brusel trpí kriminalitou nelegálních přistěhovalců a drogově závislých," tvrdí RTL. Podobné zprávy vycházející z policejních statistik jsou běžné, ale o násilí příliš nevypovídají.

Můžeme věřit tomu, že v rámci Evropské unie data nikdo záměrně nezkresluje. Ale policisté umí vyjít vstříc systému. Pokud je například jejich osobní hodnocení vázané na počet zásahů, mají motivaci hlásit co nejvíc jednoduchých případů nevyžadujících vyšetřování. Například kouření marihuany. Tím se vysvětluje, proč ve Francii za posledních 15 let statisticky vzato čtyřikrát přibylo trestných činů spojených s drogami, ačkoliv jejich spotřeba zůstává zhruba konstantní.

Co můžete udělat

Kdykoliv pochybujete o důvěryhodnosti svých dat, ověřte si je, jako by šlo o citaci nějakého politika. V příkladu s Uzbekistánem stačí zavolat někomu, kdo v zemi delší dobu žije: „Máš dojem, že je země třikrát bohatší než v roce 1995, jak tvrdí oficiální čísla?"

Co se týká policejních dat, sociologové často dělají studie, ve kterých se respondentů ptají, jestli byli terčem zločinu. Tyto studie jsou mnohem spolehlivější než policejní data. Možná proto se většinou nedostanou na titulku.

Existují i další testy, které vám pomohou lépe odhadnout důvěryhodnost dat (například Benfordův zákon), ale žádný z nich nenahradí vaše vlastní kritické myšlení.

Co přesně data říkají?

Noční práce zdvojnásobuje riziko roztroušené sklerózy

Každý duševně zdravý Němec by po přečtení tohoto titulku jistě začal odmítat noční směny. Z článku ale nevyplývá, jak velké je vlastně výsledné riziko.

Vezměte si tisícovku Němců. Roztroušená skleróza se v průběhu života objeví u jednoho z nich. Kdyby všech tisíc pracovalo v noci, počet nemocných by poskočil na dva. Noční směny tedy představují dodatečné riziko jedna ku tisíci, nikoliv sto procent. Taková informace je pro praktické rozhodování o konkrétní pracovní nabídce jistě mnohem užitečnější.

Mezi každými 15 Evropany je průměrně jeden negramotný

Výše uvedený titulek vypadá hrozivě. A je naprosto pravdivý. Mezi půl miliardou Evropanů je 36 miliónů těch, kteří neumí číst. A podle Eurostatu také 36 miliónů těch, kterým ještě nebylo sedm let.

Kdykoliv pracujete s průměrem, ujasněte si, z čeho se počítá. Je referenční populace rozdělená rovnoměrně? Díky nerovnoměrnému rozložení například většina lidí nadprůměrně dobře řídí auto. Mnozí řidiči se celý život obejdou bez nehody, případně bourají jen jednou. Naproti tomu menší počet nezodpovědných řidičů boural často, čímž tlačí aritmetický průměr nehodovosti mnohem výš, než by běžný řidič ze své zkušenosti čekal. Totéž platí o rozdělení příjmů: většina lidí má podprůměrný plat.

Co můžete udělat

Vždy berte v úvahu rozložení ukazatele v běžném vzorku. Zkontrolujte si průměr, medián i modus (nejčastěji zastoupenou hodnotu), uděláte si o datech lepší představu. Uvědomte si kontext, v jakých řádech se pohybujete; viz příklad s roztroušenou sklerózou. Konkrétní příklady poměrů („jeden ze sta“) bývají pro čtenáře výrazně srozumitelnější než procenta (1 %).

Jak spolehlivá jsou vaše data?

Problematická velikost vzorku

„80 % nespokojených se soudním systémem,“ píše španělský list Diario de Navarra. Jak ale může zobecnit výsledky od osmi set respondentů na 46 miliónů Španělů? To je ukázkové mlácení prázdné slámy. Nebo ne?

Ve skutečnosti platí, že při průzkumu velké skupiny lidí (řekněme přes několik tisíc) jen zřídka potřebujete více než tisíc respondentů, abyste dosáhli statistické chyby pod 3 %. Jinými slovy, kdybyste zopakovali průzkum s úplně jiným vzorkem, v devíti případech z deseti byste se dostali nejvýš na tři procenta daleko od výsledků z prvního pokusu. Statistika je mocná zbraň, a pokud je nějaká studie špatná, jen výjimečně je to kvůli velikosti vzorku.

Pití čaje snižuje riziko infarktu

Články o zdravotních výhodách pití čaje jsou k vidění běžně. Výjimkou není ani tento krátký článek z Die Welt, ve kterém se dočtete, že čaj snižuje riziko infarktu myokardu. Zdravotním účinkům čaje se věnuje i řada seriózních studií, ale mnohdy se zapomíná započítat vliv životního stylu – například jídelníček, povolání nebo sportovní aktivity.

Ve většině západních zemí je čaj nápojem pro vyšší třídy, které si hlídají zdravý životní styl. Pokud tedy čajové studie nezapočítají vliv životního stylu, neříkají nám o moc víc, než že bohatí lidé jsou zdravější (a nejspíš mají rádi čaj).

Co můžete udělat

Z pozice novináře nemá příliš smysl zpochybňovat číselné výsledky studie, například velikost vzorku, ledaže byste měli vážné pochyby. Vcelku snadno ale můžete zjistit, jestli autoři studie nezapomněli na nějaké zásadní relevantní informace, například korelaci pití čaje a sportování.

3.

tipy pro práci s čísly

Hledejte příběh

Abyste přitáhli čtenáře, musíte je umět přetáhnout po hlavě nějakým zásadním titulkovým číslem, které je posadí do židle a donutí přečíst zbytek článku. Příkladem takového přístupu je projekt britské novinářské neziskovky Bureau of Investigative Journalism zaměřený na Evropskou komisi a její systém finanční transparentnosti.

Autoři v databázi systému hledali konkrétní klíčová slova jako koktejl, golf nebo výjezd, aby zjistili, kolik komise utratila za příslušné položky. Výsledkem byla řada otázek a potenciálně zajímavých příběhů.

Jen s klíčovými slovy si ale člověk nevystačí. Občas se musíte zamyslet nad tím, co vlastně hledáte. V rámci téhož projektu chtěli autoři zjistit, kolik komise utrácí za soukromá letadla. Klíčové spojení „soukromé letadlo“ ale v databázi pochopitelně chybělo, a tak bylo potřeba zjistit název konkrétního dopravce („Abelag“) a vypsat z databáze výdaje za jeho služby.

Další snadný zdroj zajímavých informací získáte tím, že se v databázi budete snažit najít něco, co by v ní rozhodně být nemělo. Příkladem je společný projekt Financial Times a Bureau of Investigative Journalism zaměřený pro změnu na Strukturální fondy EU. Autoři projektu při prohledávání databáze vyšli přímo z pravidel Evropské komise, která říkají, jaký typ firem by ze strukturálních fondů žádné dotace dostávat neměl. Do této skupiny patří například výrobci tabáku, jenže v databázi fondů se přes názvy tabákových firem podařilo najít investici 1,5 miliónu eur do německé továrny firmy British American Tobacco.

Nikdy nevíte, co v databázi najdete; prostě to zkuste.

Vnímejte kontext

Nejlepší otázky jsou ty nejstarší: Je tohle opravdu velké číslo? Kde jsme ho vzali? Opravdu má takovou váhu? Obecně jde o to, abyste se naučili vnímat data jako celek, nepřehlíželi pro samé stromy les, stručně řečeno vnímali kontext.

Pokud například místní úřady po celé republice loni utratily x miliónů za kancelářské sponky, je to hodně, nebo málo? K odpovědi potřebujete kontext, který se dá získat různě. Například zdůrazněním poměru („utratili za sponky dvě třetiny svého rozpočtu na kancelářské potřeby“), vnitřním srovnáním („utratili za sponky víc než za rozvoz jídel pro seniory“) nebo vnějším srovnáním („dali loni za sponky dvakrát víc, než celý stát na mezinárodní pomoc“).

Nabízí se i další perspektiva, například vývoj v čase („rozpočet na sponky vzrostl za poslední čtyři roky trojnásobně“). Nebo můžete sestavit žebříček podle regionů či úřadů. V tom případě ovšem pozor, aby vaše srovnání bylo férové, tedy bralo v úvahu například velikost místní populace: „V přepočtu na jednoho úředníka utratí ušovický městský úřad za sponky čtyřikrát víc, než dělá republikový průměr.“

Také můžete data rozdělit na kategorie („úřady řízené stranou X utratí za sponky o polovinu víc, než úřady obsazené stranou Y“), případně zdůraznit souvislosti: „Úřady řízené politiky, kteří dostali dary od výrobců kancelářských potřeb, utrácí za kancelářské sponky víc, přičemž každá darovaná koruna se na výdajích projeví zvýšením průměrně o 100 Kč.“ Zde ovšem pozor na rozdíl mezi korelací a kauzalitou.

Užívejte si

Čísla se občas tváří nepřístupně, ale když se jimi necháte zastrašit, nikam se nedostanete. Nebojte se s nimi pohrát, prozkoumat je do hloubky. Často vás pak překvapí, jak snadno z nich dostanete nějaké tajemství nebo příběh. Prostě k nim přistupujte jako ke všem ostatním zdrojům, beze strachu a bez přehnaných očekávání. Berte práci s daty jako cvičení pro svou fantazii. Když narazíte na zjevně veliké nebo jinak nepatřičné číslo, zkuste vymyslet alternativní vysvětlení, které by mohlo lépe odpovídat datům, a ověřte si ho na dalších podkladech.

Skepse ano, cynismus ne

K datům přistupujte skepticky, ne cynicky. Zdravá nedůvěra je dobrá, cynismus znamená rozhodit rukama a vzdát se. Jestli vám datová novinařina připadá jako dobrý nápad (a jinak byste tenhle text nečetli), musíte přistoupit na to, že data jsou něco mnohem víc než příslovečné lži

a zatracené lži nebo pouhý odrazový můstek k atraktivním a zavádějícím titulkům. Správně zpracovaná data jsou zásadní zdroj informací. Nesmíme být ani cyničtí, ani naivní, ale pozorní.

Nejdřív data, potom závěry

Když vám někdo řekne, že se během hospodářské recese hodně pije, usoudíte, že je to všeobecnou depresí. Když vám poví, že se během recese pije málo, pomyslíte si, že jsou všichni bez peněz. Jinými slovy: ať už data tvrdí cokoliv, vy máte předem připravenou interpretaci, že jde všechno od desíti k pěti. Když se pije hodně, je to špatně. Když se pije málo, je to špatně. Pokud máte pracovat s daty, musíte je nechat mluvit a neválcovat je vlastními náladami, názory a hypotézami. V dnešní době je k dispozici tolik dat, že se při troše snahy dá potvrdit prakticky cokoliv. Datová žurnalistika nepřináší žádnou podstatnou hodnotu, pokud ji neděláte s otevřenou hlavou. Pokud má být objektivní, musíte se o to postarat vy. Čísla nejsou objektivní sama od sebe.

Nejistota není sprosté slovo

Zvykli jsme si čísla spojovat s autoritou a jistotou. Přitom se ale běžně stává, že naše nejlepší odpověď zní nevím. Nebo je tak nepřesná, že bychom se s ní radši vůbec neukazovali na veřejnosti. Takové věci je potřeba říkat nahlas. Možná vám to připadne jako dobrý způsob, jak torpédovat svůj vlastní článek. On to ale naopak může být i dobrý způsob, jak přijít na nové otázky. Často se také stává, že data jdou vyložit dvěma různými legitimními způsoby. Čísla nejsou nutně černobílá.

Psaní je příběh

Příběh o vašem pátrání, o postupu od jednoho důkazu ke druhému, může posloužit jako skvělá kostra článku. Dvojnásob to platí v datové žurnalistice, kde si jen zřídka vystačíte s jedním číslem. Nové zdroje přináší nové úhly pohledu, nové nápady, lepší celkový obrázek. Nesnažte se nutně působit jako autorita, která čtenářům naservíruje až hotovou odpověď.

4.

ako získavať informácie

Táto kapitola opisuje najčastejšie spôsoby získavania informácií, ktoré používajú investigatívni a dátoví reportéri. Tvrdenia, ktoré v texte nájdete, vychádzajú iba zo skúseností a znalostí autora. Niektoré z tvrdení môžu časom stratiť aktuálnosť, pretože budú spochybnené novými rozhodnutiami slovenských súdov.

Najprv hľadajte

Kým sa začne novinár oháňať zákonom č. 211/2000 Z. z. o slobodnom prístupe k informáciám, takzvaným infozákonom, mal by sa na dáta a informácie pýtať a hľadať ich. Na úradných weboch je totiž množstvo dát, iba sú schované a nie sú dostupné z jedného centrálneho katalógu datasetov.

Príkladom sú štatistiky o vývoze a dovoze zbraní na webe ministerstva hospodárstva, na základe ktorých sa dajú robiť prehľadné vizualizácie o tom, kde sa strieľa z legálne vyvezených slovenských zbraní.

Prístupné sú aj informácie o kontrolách obchodných reťazcov (na webe Štátnej veterinárnej a potravinovej správy), ktoré po analýze ponúknu obraz o tom, kde sú najčastejšie pokutované obchody alebo za čo dostávajú podnikatelia najvyššie pokuty.

Podobne podrobné dáta o pohyboch lietadiel nad Slovenskom, ktoré sú na webe slovenského riadenia letovej prevádzky. Preto treba najprv hľadať, či dáta, ktoré chcete, už niekde nevisia.

Veľmi dobrým pomocníkom je Datanest Aliancie Fair-play – rozsiahly katalóg rôznych údajov, ktoré úrady zverejnili na rôznych miestach na internete. Bonusom sú datasety, ktoré Aliancia Fair-play získala, očistila, spracovala a zverejnila.

Pýtajte sa

V prípade, že novinár nenájde na webe údaje, o ktoré sa zaujíma, je ešte stále použitie infozákona predčasné. Treba radšej zavolať na úrad, o ktorom sa domnieva, že údajmi disponuje.

Najprv sa treba obrátiť na kompetentný odbor – ministerstvá aj úrady majú zvyčajne na weboch menný zoznam, preto keď sa zaujímate napríklad o register lodí, treba na Dopravnom úrade volať rovno na odbor vnútrozemskej plavby.

Títo ľudia zvyčajne vedia, v akom stave je dataset a či vôbec a ako sa môže zverejniť.

V takomto prípade vás najčastejšie kompetentní ľudia presmerujú na hovorcu úradu alebo ministerstva, aby ste tému riešili s ním. To predĺži celý proces získavania informácií, lebo hovorca je len sprostredkovateľom medzi novinármi a kompetentnými ľuďmi, ktorí žiadaný dataset spravujú.

Komunikácia s hovorcami je niekedy zložitá, pretože podľa zákona nemá žiadne pravidlá. Tlačový zákon hovorí iba toľko, že verejné inštitúcie „sú povinné na základe rovnosti poskytovať (novinárom) informácie o svojej činnosti na pravdivé, včasné a všestranné informovanie verejnosti“.

V preklade to znamená, že keď sa hovorca rozhodne, že žiadny dataset nedostanete, tak žiadny dataset nedostanete a nezmeníte to. Napriek tomu, že získať datasety cez dohodu s úradom sa podarí málokedy, novinári by sa o to mali pokúšať, pretože v konečnom dôsledku je to najrýchlejšie a najľahšie.

Čo vlastne chcem?

Keď novinár zistí neochotu zo strany úradov dobrovoľne vydať dataset alebo akékoľvek iné informácie, je načase použiť postup podľa zmieneného infozákona.

V prvej fáze, keď iba posielate žiadosť o informácie, je to ľahké – najviac času treba venovať sformulovaniu toho, čo vlastne chcete získať. Pri spisovaní žiadosti treba myslieť na limity infozákona. Napríklad podľa rozhodnutia Najvyššieho súdu úradníci nemusia vytvárať nové informácie.

A práve otázku, čo je ešte prepisovanie údajov a čo už je vytváranie nových informácií, často riešia úradníci pri sprístupňovaní datasetov.

Preto sa môže stať, že keď si vypýtate informáciu o počte stoličiek podľa farebnosti na danom úrade, nemusíte ju dostať, lebo úradníci nie sú povinní ich spočítať. Treba si vypýtať ich zoznam z poslednej inventarizácie a z neho si to sami spočítate. Platí to, samozrejme, aj na iné dáta.

Podobne sa dajú získať aj rôzne metadáta o činnosti úradov. Niekedy treba dať úradníkom slušne najavo, že ste ochotný naskočiť na ich byrokratický spôsob myslenia a budete sa s nimi o informácie sporiť. Občas sa stane, že ustúpia, dáta poskytnú a ušetria si tak zbytočnú administratívu okolo.

Pri písaní žiadosti o informácie si preto treba dávať pozor najmä na formuláciu požiadavky, tá je totiž niekedy polovicou úspechu. Na webe nájdete množstvo návodov, aké formálne podmienky musí taká žiadosť spĺňať.

Treba tam najmä napísať meno, adresu, čo žiadate, ako vám to majú poslať (elektronicky, papierovou poštou), napísať aj e-mail.

Na vybavenie žiadosti majú úradníci osem pracovných dní, lehotu môžu predĺžiť najviac o ďalších osem pracovných dní. Vo vývoji je aj projekt ChcemVediet.sk, cez ktorý si budete môcť napísať žiadosť bez toho, aby ste poznali základné pravidlá. Celý proces vybavovania žiadosti vrátane odpovedí úradu by potom mal byť zverejnený na webe. Na vytvorenie žiadosti o sprístupnenie akéhokoľvek súdneho rozhodnutia sme zas s programátorom Matejom Lukášikom vytvorili projekt nezverejnili.sk.

- „Zákon o slobodnom prístupe k informáciám v žiadnom ustanovení neukladá povinnému subjektu spracovávať údaje do určitých databáz.“ (rozsudok Najvyššieho súdu SR vo veci sp. zn. 4Sžso/31/2008)

- „O vytvorenie novej informácie nejde, ak na prípravu odpovede na žiadosť o informácie stačí iba mechanické vyhľadanie a zhromaždenie údajov, ktoré sa u povinnej osoby nachádzajú a ktoré musia byť napríklad vyhľadané v rôznych dokumentoch, ‚vyňaté‘ z týchto dokumentov, zhromaždené a následné ‚vtelené‘ do odpovedi na žiadosť.“ (Peter Wilfling v publikácii Zákon o slobodnom prístupe k informáciám Komentár)

Samostatnou kapitolou sú žiadosti o informácie, ktoré novinári chcú získať od Európskej únie. Na Slovensku ide ešte stále o ojedinelú, no rozširujúcu sa situáciu, pretože Komisia má k dispozícii množstvo štatistických súhrnov a zároveň veľa zaujímavých dokumentov pre investigatívnych žurnalistov (napríklad z rôznych vyšetrovaní).

Na Európsku komisiu neplatí slovenský infozákon, ale nariadenie Európskeho parlamentu a Rady (ES) č. 1049/2001 z 30. mája 2001 o prístupe verejnosti k dokumentom Európskeho parlamentu (EP), Rady a Komisie.

Praktický postup je veľmi podobný žiadosti podľa slovenského infozákona: musíte napísať, čo chcete, od ktorej inštitúcie, v akej forme to žiadate sprístupniť (elektronicky alebo poštou), žiadosť stačí poslať e-mailom.

O prijatí žiadosti dostanete potvrdenie a orgán EÚ, ktorý ste o dáta alebo informácie žiadali, má potom 15 pracovných dní, aby žiadosť posúdil a vyrozumel vás, či vám informácie sprístupní. Komunikovať sa dá aj v slovenčine, ale angličtina celý proces zrýchľuje.

Európska komisia má navyše zriadený Register dokumentov Komisie, kde sama dokumenty proaktívne zverejňuje. O dokument, ktorý tam nenájdete, zas môžete žiadať na tejto adrese (žiadosť cez tento web nahrádza žiadosť e-mailom). Žiadosť podľa nariadenia č. 1049/2001 sa dá navyše podať aj cez web AskTheEU.org (po anglicky, nemecky, španielsky, francúzsky) – čo je aplikácia na jednoduché podanie žiadosti o informácie podľa nariadenia č. 1049/2001.

Vojna paragrafov

V prípade, že žiadosť o informácie podľa slovenského infozákona nebola vybavená do ôsmich pracovných dní alebo žiadosť zamietli (informácie neboli sprístupnené alebo boli sprístupnené iba sčasti), môžete sa odvolať v lehote do 15 kalendárnych dní odo dňa doručenia rozhodnutia o nesprístupnení informácií.

Keď vám žiadne rozhodnutie neprišlo a úradníci vašu žiadosť úplne odignorovali, odvolanie sa podáva proti fiktívnemu rozhodnutiu – zákon predpokladá, že takéto fiktívne rozhodnutie sa vydá v ôsmy pracovný deň od podania žiadosti a doručí na jedenásty deň od podania žiadosti, od tohto dátumu sa potom počíta 15-dňová lehota na podanie odvolania.

Spísanie samotného odvolania je veda. Základom je, aby ste v odvolaní napísali, aké rozhodnutie napádate, čo mu vyčítate a ako žiadate, aby odvolací orgán rozhodol.

Praktickú príručku, ako žiadať informácie a ako sa odvolať, ktorú mám stále po ruke, napísal spolupracovník združenia VIA IURIS Peter Wilfling a môžete si ju stiahnuť ako PDF.

VIA IURIS má návod, tipy & triky ohľadom infozákona aj na svojom webe. Investigatívnym reportérom odporúčam kúpiť si aj Judikatúru vo veciach slobodného prístupu k informáciám, ktorú zostavil sudca Najvyššieho súdu SR Ivan Rumana spolu s asistentkou senátu Najvyššieho súdu SR Inou Šingliarovou.

Pomocníkom v paragrafovej vojne pri žiadaní o informácie môže byť aj Přehled judikatry vo věcech práva na informace od českého advokáta Františka Korbela. Vybrané rozhodnutia českých súdov síce nie sú priamo použiteľné na Slovensku, no dajú sa použiť ako podporná argumentácia v odvolaniach.

O odvolaní proti rozhodnutiu rozhoduje nadriadený správny orgán – napríklad keď žiadate informácie od ministerstva, proti odvolaniu rozhoduje minister, keď žiadate informácie od okresného súdu, rozhoduje o odvolaní krajský súd, keď od krajského súdu, odvolanie príde na stôl ministerstvu spravodlivosti, keď žiadate informácie od akciovej spoločnosti, ktorú vlastní štát alebo samospráva, o odvolaní rozhodne jej zakladateľ (teda mesto alebo ministerstvo, ktoré spravuje podiel vo firme).

Odvolanie sa podáva písomne alebo elektronicky so zaručeným elektronickým podpisom cez dátovú schránku na Slovensko.sk tej istej inštitúcie, ktorej ste prvotne poslali žiadosť o informácie.

Má 15 dní na to, aby sama zvážila, či odvolaniu vyhovie, alebo ho posunie nadriadeným, a nadriadený orgán má ďalších 30 dní na rozhodnutie o odvolaní. Inak povedané, na to, čo sa stane s vaším odvolaním, môžete čakať aj 45 dní.

Platí preto, že právna bitka o informácie sa vypláca najmä investigatívnym reportérom, ktorým ide skôr o obsah informácie, než o rýchlosť jej získania. Pri aktuálnych témach, kde sa vyžaduje rýchle spracovanie témy, je totiž dvoj- či trojmesačná korešpondencia s úradmi nemysliteľná.

V prípade, ak vám nadriadený orgán odpíše, že odvolaniu nevyhovel, môžete toto rozhodnutie žalovať. Tento postup odporúčam, no treba si dohodnúť advokáta (sami žalobu podať nemôžete, pred súdom musíte byť zastúpený advokátom) a 70 eur sa platí ako správny poplatok za podanie žaloby.

Podobný postup funguje aj v prípadoch, keď neúspešne žiadate o informácie od európskych inštitúcií podľa nariadenia č. 1049/2001. Keď od nich dostanete odpoveď, že žiadané informácie nesprístupnia, môžete do 15 pracovných dní podať opakovanú žiadosť (akási obdoba odvolania), kde treba podrobne vysvetliť, prečo si myslíte, že utajenie informácií nie je správne.

Opakovaná žiadosť sa dá podať elektronicky, netreba ju posielať papierovou poštou. Európska inštitúcia má 15 pracovných dní, aby na ňu reagovala. Keď ani na opakovaný pokus informácie nesprístupní, môžete podať sťažnosť európskemu ombudsmanovi (netreba byť zastúpený advokátom) alebo žalobu na Súdny dvor EÚ (treba byť zastúpený advokátom).

[redacted]
adam.valcek@gmail.com

5.

extrakce dat z webu

Ideální je, když se vám data podaří na webu najít v nějakém přímo zpracovatelném formátu, například jako excelovou tabulku nebo ve formátu CSV. Občas se ale stane, že data na webu najdete, ale nejsou ke stažení v rozumném formátu a obyčejné kopírování přes schránku nepřipadá v úvahu nebo nefunguje. Nemusíte propadat panice, ještě existuje několik možností:

Stahování dat prostřednictvím API. Moderní webové služby, například online databáze a sociální sítě (včetně Twitteru, Facebooku a dalších) dnes kromě běžného uživatelského rozhraní často nabízí také API neboli application programming interface, rozhraní určené strojům. To je fantastický způsob, jak se dostat k vládním i komerčním datům, včetně informací ze sociálních médií.

Extrakce dat z PDF. Značně pracná varianta, protože PDF je formát určený primárně pro popis tištěné stránky a neuchovává všechny informace o struktuře dat, která jsou v dokumentu uložena. Konkrétní návod je mimo rozsah této publikace; nástroje a tipy pro extrakci dat z PDF najdete na webu.

Screen scraping neboli extrakce dat přímo z webových stránek. U této varianty vyzobáváte informace prostřednictvím speciálního programu nebo vlastního kusu kódu z webové stránky, která nebyla primárně určena pro strojové zpracování. Scraping je velice silný nástroj a dá se použít téměř všude, ale vyžaduje určité technické znalosti webu.

Přes všechny pěkné technické varianty nezapomínejte na jednoduchá řešení: často se vyplatí ještě chvilku hledat soubor se strojově čitelnými daty nebo prostě zavolat instituci, jejíž data potřebujete. A pokud nic z toho nevyjde, můžete se pustit do scrapování, nad kterým se teď na chvíli zastavíme.

Co jsou strojově čitelná data

Když hledáte data pro další zpracování, vaším cílem jsou většinou strojově čitelná data. Což znamená data uložená s ohledem na další automatické zpracování počítačem, nikoliv prezentaci lidem; data strukturovaná podle logiky uložených informací, nikoliv podle budoucího zobrazení. Mezi strojově snadno čitelné formáty patří například CSV, XML, JSON nebo excelové tabulky. Naopak dokumenty z textových procesorů (Word a podobně), soubory ve formátu PDF a do jisté míry také HTML soubory se zabývají spíše vizuálním rozložením informací. Zejména formát PDF byl původně určen pro komunikaci s tiskárnou, takže pracuje spíše s umístěním jednotlivých čar a teček na stránce, nikoliv s vyššími celky jako písmeny, slovy, odstavci, tabulkami a podobně.

K čemu je scrapování

Určitě jste to zažili sami: najdete na webu zajímavou tabulku a zkusíte si ji zkopírovat do Excelu, abyste ji mohli nějak zpracovat nebo uložit na později. Jenže to v praxi často nefunguje, případně jsou informace roztroušené do mnoha samostatných stránek. Ruční kopírování rychle omrzí, takže má smysl místo něj použít kus kódu, který práci udělá za vás.

Velká výhoda scrapování je v tom, že se dá použít prakticky u jakéhokoliv webu, od předpovědi počasí po přehled vládních výdajů, a to i když server nenabízí API pro přístup ke strojově čitelným datům. I scrapování ale pochopitelně má své limity. Automatická extrakce dat je složitější, neprakticky náročná nebo rovnou nemožná například v následujících případech:

- Stránky se špatným HTML kódem, který poskytuje jen minimum informací o struktuře dokumentu. Klasickým příkladem jsou starší vládní weby.
- Systémy přímo stavěné proti automatickému zpracování, například CAPTCHA nebo paywally, platební zdi umožňující přístup pouze platícím uživatelům.
- Weby, které spoléhají na funkce interaktivního webového prohlížeče, například JavaScript nebo cookies.
- Weby, na kterých chybí úplné seznamy i možnost vyhledávat, takže se při scrapování nemáte od čeho odrazit a museli byste ručně procházet jednu stránku po druhé.
- Zákaz automatického zpracování ze strany správců serveru.

Problematická může být i právní stránka věci; právní systém některých zemí omezuje možnosti nakládat s daty publikovanými online. Jako novinář v tomto ohledu můžete a nemusíte mít zvláštní práva. Scraping veřejně dostupných vládních dat by měl být bezproblémový,

jen se dvakrát ujistěte, než data budete publikovat. Komerční organizace a některé neziskovky bývají méně tolerantní a protože scraping může nadměrně zatěžovat jejich server, v krajním případě ho mohou vnímat jako DDoS útok. Stažené informace se také mohou týkat soukromí osob, takže byste mohli mít problémy se zákony na ochranu osobních údajů nebo profesními etickými kodexy.

Scrapovací nástroje

Programů, které se dají použít pro extrakci informací z webových stránek, existuje široké spektrum, od online služeb po rozšíření webového prohlížeče. Služba Readability vám například pomůže vytáhnout z webové stránky čistý text, rozšíření DownThemAll pro Firefox usnadňuje stahování většího počtu souborů a rozšíření Scraper pro Google Chrome je přímo stavěné na kopírování tabulek z webových stránek.

Praktické jsou také funkce prohlížečů určené vývojářům. Díky nim se můžete podívat, jak je stránka strukturovaná a co si váš prohlížeč povídá se serverem na druhé straně. Google Chrome, Safari a Internet Explorer mají vývojářské nástroje vestavěné, pro Firefox si můžete stáhnout rozšíření FireBug.

Přímo na scraping se specializuje server ScraperWiki, kde si můžete snadno napsat scraper v Pythonu, Ruby nebo PHP. Je to ideální způsob, jak začít se scrapováním, aniž byste se museli mořit s instalací vývojářských nástrojů na svůj vlastní počítač. Scrapování do určité míry podporují i další rozšířené webové služby, například Google Docs nebo Yahoo! Pipes.

Technické principy scrapování

Klikací nástroje zmíněné v předchozím oddílu jsou výborný začátek, ale dříve nebo později se většinou budete muset ponořit do scrapovaných stránek a najít, kde přesně se v nich hledané informace nachází. Nejde o žádné velké programování, jen musíte mít základní představu o struktuře webových stránek a databáze, ze které těžíte.

HTML stránka se uvnitř skládá z mnoha takzvaných tagů neboli značek, které strukturují holý text stránky do větších logických celků (například odstavců, tabulek nebo odkazů) a vkládají do něj další objekty, například obrázky. Ke značkám mohou být pomocí takzvaných atributů připojené další informace. Často mívá značka například jedinečný identifikátor, podle kterého ji můžete snadno najít v celém dokumentu. Běžné je také podrobnější rozdělení značek jednoho typu do několika různých tříd.

Všechny tyto jazykové nástroje mají jediný cíl: vnést do textu stránky strukturu, aby se dal snadno formátovat a zpracovávat. A právě toho se využívá i při scrapování dat. Nejprve si prohlédnete zdrojový kód stránky (například pomocí vývojářských doplňků prohlížeče), abyste zjistili, kde přesně se ve změti značek nachází potřebné informace. Pak napíšete malý program, takzvaný scraper, který podle vašich instrukcí sáhne na ta správná místa v dokumentu a data vytáhne.

Příklady scraperů, ze kterých se můžete odrazit při vlastním experimentování, najdete na zmiňovaném serveru ScraperWiki.

6.

tipy na prácu s internetovými zdrojmi

WHOIS

je stručne povedané register vlastníkov domén, IP adries a ďalších internetových objektov. V poslednom čase vlastníci domén často používajú takzvanú súkromnú registráciu, pri ktorej sa v registri neobjavia ich osobné údaje, iba názov organizácie, cez ktorú doménu registrovali. V mnohých ostatných prípadoch môžete podľa názvu domény v systéme WHOIS zistiť meno vlastníka, jeho adresu, e-mail aj telefónne číslo, ale aj IP adresu, podľa ktorej dospejete k informáciám o jednotlivcovi alebo o organizácii, ktorej počítač alebo server s touto adresou patrí. To sa výborne hodí napríklad vtedy, keď sa snažíte zistiť identitu používateľa webovej služby, pretože väčšina serverov si IP adresu svojich návštevníkov archivuje. www.whois.icann.org

Operátori vo vyhľadávači Googlu

Pri vyhľadávaní cez Google sa dajú používať rôzne typy operátorov. Napríklad keď pridáte k svojej otázke reťazec site:.sk, Google vráti iba výsledky z adries, ktoré používajú slovenskú doménu. Dá sa to využiť napríklad pri hľadaní na vládnych stránkach (site:.gov.sk) alebo pri hľadaní v zahraničných zdrojoch. Výsledky môžete potom zúžiť aj na konkrétne podadresy, napríklad site:domena.cz/dokumenty. Tento trik je zvlášť praktický pri hľadaní obsahu, ktorý vlastník internetovej stránky síce zverejnil, ale nehrnie sa do jeho propagácie. Ďalším užitočným operátorom je filetyp:, ktorý slúži na vyhľadávanie konkrétneho typu súborov, napríklad pre PDF je operátor filetyp:pdf. Existujú aj ďalší operátori, ich zoznam je napríklad tu. Dajú sa rôzne kombinovať.

Google Cache

Kontroverzné stránky môže ich autor bez varovania stiahnuť alebo zmeniť. Ak sa potrebujete dostať k pôvodnému zneniu, najprv vyskúšajte verziu stránky v medzipamäti vyhľadávača Google – ide o podobu stránky, ako si ju Google zapamätal pri poslednom indexovaní. Vyhľadávač si pamätá vždy iba poslednú verziu, takže treba hľadať rýchle, kým Google nezaindexuje aj novú verziu stránky už so zmeneným obsahom. Zadajte do Google ako vyhľadávaciu otázku URL stránky a keď sa vám objavia výsledky, nájdete si pri odkaze link na cache (slovenský Google používa pojem v pamäti). Ak nájdete v medzipamäti to, čo hľadáte, urobte si printscreen alebo si inak skopírujte archivovanú stránku; cache Googlu sa totiž môže opäť aktualizovať.

Kataster nehnuteľností

Úrad geodézie, kartografie a katastra SR už dlhšie testuje novú interaktívnu katastrálnu mapu – Mapka, ktorá funguje v najpoužívanejších prehliadačoch. Netreba inštalovať žiadne špeciálne prídavky, ako si to vyžaduje prehliadanie mapy na katasterportal.sk. Mapka navyše zobrazuje katastrálnu mapu spolu s tradičnými údajmi, ako sú názvy úloh, čísla domov, dá sa zobraziť aj prekrytie katastrálnej mapy a Google Maps tak, aby ste mali presnú predstavu, kde sa nehnuteľnosť naozaj nachádza. Po kliknutí na parcelu sa zobrazia na ľavom paneli základné informácie o majiteľoch a liste vlastníctva. Na získanie bližších informácií o majiteľoch potom slúži namiesto katasterportal.sk projekt cica.vugk.sk, kde môžete vlastníkov či listy vlastníctva vyhľadávať bez časového obmedzenia a nutnosti neustále vypĺňať captch.

Obchodné registre, akcionári a účtovné výkazy

Okrem základného vyhľadávanie cez web orsr.sk sa dá obchodný register pohodlnejšie prehliadať aj pomocou spomínaných Google operátorov, napríklad cez site:orsr.sk kľúčové slovo. Tento spôsob vyhľadávania je v mnohých prípadoch pohodlnejší, keďže vyhľadávanie priamo cez vyhľadávacie formuláre na orsr.sk nie je vždy presné. Pri hľadaní majiteľov akciových spoločností, ktoré nemajú akcionárov zapísaných priamo na orsr.sk, sa dá hľadať v účtovných výkazoch v účtovnej uzávierke. Tie sa zverejňujú na webe registeruz.sk v nespracovanej podobe, v poznámkach k účtovnej uzávierke sa potom zvyknú uvádzať mená spoločníkov, resp. akcionárov spoločnosti k poslednému dňu daného účtovného obdobia. V účtovných výkazoch sa navyše dajú nájsť on-line aj ďalšie zaujímavé údaje, napríklad o pôžičkách spriazneným osobám (vlastníkom, manažérom, zamestnancom a pod.). Obraz o vzťahoch v pozadí biznisu sa dá cez internet dotvoriť aj vďaka Notárskemu centrálnemu registru záložných práv, ktorý ponúka zápis záložných práv, s opismi záloh a často aj úverových vzťahov. V registri tak napríklad nájdete, aká firma si od akej banky na čo vzala úver a čím za tento úver ručí.

Archív webu

Zmeny konkrétneho serveru alebo stránky za dlhšie časové obdobie, povedzme, mesiace alebo roky, si môžete pozrieť pomocou služby Wayback Machine, ktorá pravidelne archivuje veľkú časť webu. Stačí zadať adresu stránky, ktorá vás zaujíma, a ak je v archíve, zobrazí sa vám kalendár s vyznačenými archivovanými verziami stránky. Po kliknutí na konkrétny deň vám archív ukáže obsah stránky, ktorý zhruba zodpovedá vtedajšej skutočnej podobe. Často chýba formátovanie alebo obrázky, ale text a základná podoba by mali ostať, čo na predstavu stačí.

Hľadanie podľa obrázka

Niekedy sa hodí vedieť, odkiaľ pochádza obrázok, ktorý máte uložený na lokálnom disku, resp. ktorý ste našli niekde na webe. Google už umožňuje vyhľadávanie cez obrázok, stačí nahrať súbor alebo skopírovať link a o ostatné sa postará Google. Múdre algoritmy porovnávajú obrázky na webe a zobrazia výsledky s podobnými alebo s rovnakými obrázkami spolu s informáciami o stránke, kde sa nachádzajú. Podobnú službu ponúka napríklad TinEye. Služba sa výborne hodí v prípadoch, keď máte podozrenie, že už ste niekde obrázok vydávaný za novinku alebo originál videli.

Google Trends

Služba ponúka obraz o tom, čo ľudia hľadajú na webe. Google Trends zverejňuje svoje štatistiky často vyhľadávaných fráz. Môžete zadať jednu konkrétnu frázu (Váhostav) alebo viac fráz naraz oddelených čiarkou (Andrej Kiska, Robert Fico) a sledovať, ako sa v čase menil dopyt ľudí po vyhľadávaní týchto slov. Výber dát sa dá zúžiť podľa rôznych kritérií (regionálne, časové a pod.). Chýbajú absolútne čísla, grafy ukazujú iba relatívny záujem o danú frázu v percentách, čo nemá vždy jasnú vypovedaciu hodnotu. Pri niektorých frázach zas Google nemá dosť zdrojových dát, a tak žiadne štatistiky nezobrazí.

7.

vizualizace jako tažný kůň datové žurnalistiky

Ještě než data začnete vynášet do grafů a map, zamyslete se na moment nad tím, jakou roli vlastně hraje interaktivní a statická grafika ve vaší práci.

Během přípravy podkladů vizualizace pomáhá hledat témata a otázky pro zbytek článku, upozorňuje na anomálie (ať už jde o chyby v datech nebo náměty na dobrý článek), pomáhá hledat typické příklady nebo ukazuje díry ve vašich zdrojích.

Uplatní se samozřejmě ale i ve výsledném článku, kde dokáže přesvědčivě ilustrovat pointu nebo zbavit text nadbytečných technických detailů. Navíc přináší do vaší práce větší transparentnost, zvlášť pokud jde o interaktivní vizualizace, ve kterých se čtenáři mohou volně pohybovat.

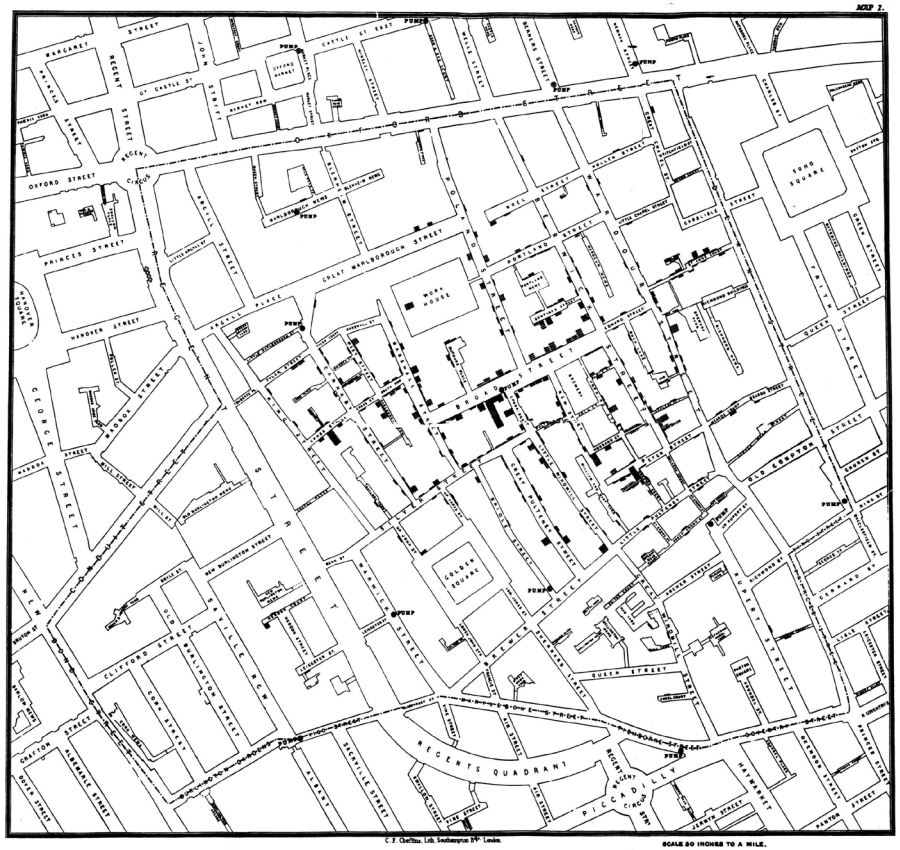
Z toho plyne, že byste s vizualizacemi měli začít záhy a průběžně je aktualizovat. Neberte vizualizaci jako samostatný krok, na který přijde čas, až bude článek z větší části hotový. Používejte vizualizaci jako další vodítko pro svou práci.

Různé grafy, různé příběhy

V dnešním digitálním světě, kde už ani trojrozměrná virtuální realita není nic neobvyklého, máme sklon zapomínat, že jsme dlouhou dobu měli k dispozici jen inkoust a papír. Statický a plochý papír dnes považujeme za médium druhé kategorie, ale faktem je, že za stovky let psaní a tisku se nám podařilo shromáždit obrovský arzenál nástrojů pro reprezentaci dat na papíře. Interaktivní grafy, vizualizace dat a infografiky, které jsou dnes ohromně v kurzu, často ignorují užitečné historické zkušenosti. Je na nás, abychom tyto zkušenosti přenesli do nových médií.

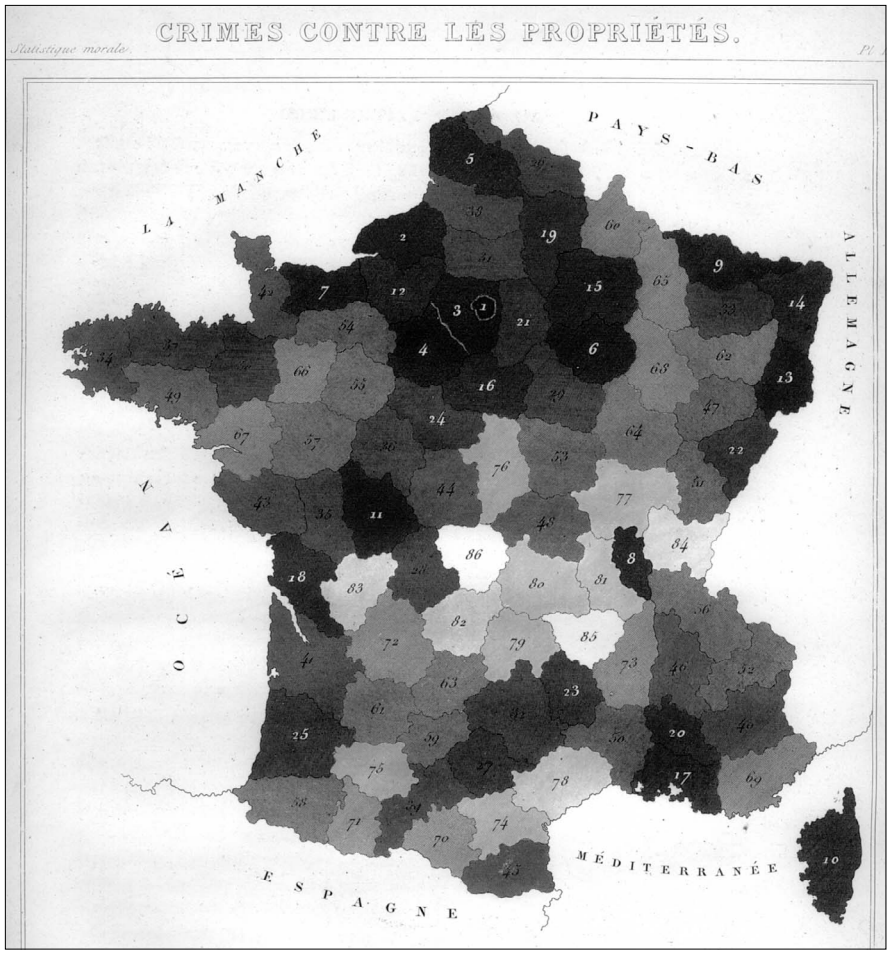
Některé z nejznámějších diagramů a grafů vzešly z potřeby přehledně popsat složitá tabulková data. To bylo také častým úkolem Williama Playfaira, skotského polyglota žijícího na přelomu 18. a 19. století, který pro svět objevil řadu grafů používaných dodnes. Například ve své knize Commercial and Political Atlas, vydané roku 1786, představil klasický sloupcový graf, kterým nově a přehledně ilustroval skotský import a export.

Následovala kniha Statistical Breviary, kterou Playfair v roce 1801 popularizoval dnes tolik obávaný „koláč“. Původním motivem pro zavádění nových diagramů a grafů byl obchod, ale s postupem času se objevovaly i další, z nichž některé přímo zachraňovaly životy. V roce 1854 John Snow vytvořil svou proslavenou mapu londýnské epidemie cholery, kde nad každou adresou s hlášeným výskytem nemoci nakreslil malý černý obdélník. Za krátký čas se černé značky jasně nakupily kolem problematické pumpy, která tím byla odhalena jako zdroj nákazy, a problém byl vyřešen. **Cholerová mapa Londýna; John Snow ↓**



V průběhu let se nový obor osmělovat ke stále odvážnějším experimentům a posouval médium až k jeho dnešní podobě. André-Michel Guerry jako první přišel s myšlenkou takzvaného choroplethu neboli mapy, na které jsou jednotlivé regiony obarvené podle nějaké proměnné; v roce 1829 vybarvil mapu Francie podle úrovní kriminality. Dnes se tyto mapy běžně používají pro popis volebních preferencí a výsledků, rozložení příjmů a řady dalších ukazatelů vázaných na zeměpisnou oblast. Nápad je to v principu velmi jednoduchý, ale pokud nemá výsledná mapa zkreslovat a má být srozumitelná pro čtenáře, vyžaduje jisté úsilí.

Kriminalita ve Francii; André-Michel Guerry ↓



Dobrý novinář by měl mít v aktivním repertoáru řadu vizualizačních nástrojů. Nemá smysl začínat těmi složitými, důležité je bezpečně zvládnout základy. Ať už budete dělat cokoliv, opírat se vždy budete o několik jednoduchých výchozích grafů a diagramů. Teprve z tohoto pevného zázemí se můžete pustit do složitějších vizualizací.

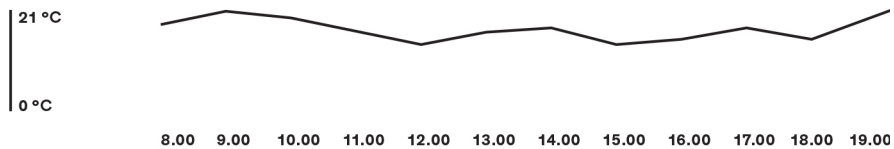
Mezi ty nejzákladnější typy grafů patří čárové a sloupcové grafy. Používají se v podobných případech, ale jsou mezi nimi i podstatné významové rozdíly.

**Jednoduchý sloupcový graf,
ideální pro reprezentaci nespojitých dat**

Podívejme se například na měsíční statistiku firemních příjmů za jeden rok. Při popisu sloupcovým grafem dostaneme 12 sloupečků, z nichž každý ukazuje zisk za jeden měsíc roku.

Mohli bychom místo sloupců použít čárový graf? Problém je v tom, že čárový graf se hodí spíše pro spojitá data. Naše čísla o příjmech firmy spojitá nejsou, ukazují součet příjmů firmy za daný měsíc. Ze sloupcového grafu vidíme, že za leden firma vydělala $100 a za únor $120. Kdybychom graf změnili na čárový, na první dny měsíce by vycházela táž čísla, ale z průběhu čáry bychom mohli získat dojem, že někdy v polovině ledna firma vydělala $110. Což není pravda. Pro nespojitá data se víc hodí sloupcový graf; čárový graf je ideální pro data spojitá, například průběh teplot.

**Jednoduchý čárový graf,
ideální pro reprezentaci spojitých dat**

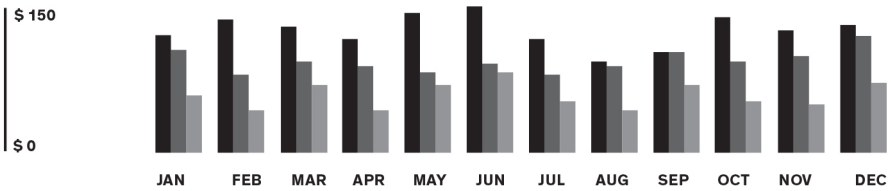


Na čárovém grafu teplot vidíme, že v osm ráno byla teplota 20 °C a o hodinu později 22 °C. Podle průběhu čáry můžeme odhadnout, že v 8.30 mohlo být kolem 21 °C. Tentokrát to dává smysl, protože průběh teploty je spojitý – jednotlivé body grafu nepopisují součet nějakých čísel, nýbrž konkrétní hodnotu v daném čase nebo její odhad mezi dvěma měřeními.

Sloupcový i čárový graf mají skupinovou variantu, kterou už se dají velmi pěkně vyprávět příběhy. Vezměme si například firmu se třemi pobočkami.

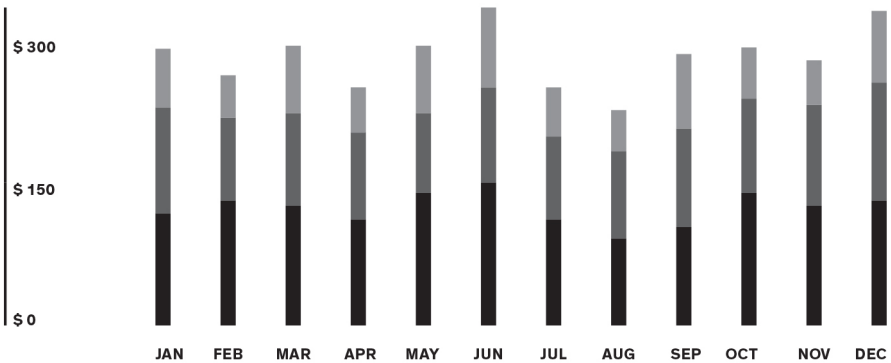
Skupinový sloupcový graf

Teď máme za každý měsíc tři sloupce, jeden pro každou pobočku, 36 celkem za jeden rok. Ve skupinovém grafu na první pohled vidíme, která pobočka byla v daném měsíci nejziskovější. To je zajímavý a legitimní úhel pohledu, ale nám se nad stejnými daty nabízí ještě druhý.



Skládaný sloupcový graf

Když sloupce naskládáme nad sebe, vznikne takzvaný skládaný sloupcový graf, ve kterém už sice tak dobře nevidíme srovnání jednotlivých poboček mezi sebou, ale zase je jasnější, ve kterém měsíci nejvíc vydělává firma jako celek.



Oba grafy dávají smysl, a přestože vychází ze stejných dat, každý mluví o něčem jiném. Pro vás jako novináře pracujícího s daty se tu nabízí zásadní otázka, kterou si musíte zodpovědět hned na začátku: O čem vlastně chcete psát? O tom, který měsíc je nejlepší k podnikání, nebo o tom, která pobočka táhne firmu? Tohle byl jen triviální příklad, který ovšem ilustruje základní princip datové žurnalistiky. Na prvním místě jsou správné otázky, teprve na druhém výpočty. Váš příběh si sám řekne, jaká vizualizace je pro něj nejlepší.

Sloupcový a čárový graf jsou denním chlebem každého datového novináře. Po jejich zvládnutí můžete svůj arzenál rozšířit o histogramy a další typy diagramů (například horizontové, sparkline nebo proudové grafy), které mají společný základ a specializují se na různé situace, ať už podle množství dat, jejich zdroje nebo vzájemného vztahu mezi textem a grafikou.

Velice často se v žurnalistice používají mapy. Většinou nás zajímá srovnání nějakého ukazatele mezi dvěma místy, tok dat z jednoho regionu do druhého a podobně. Klíčová otázka zní, jak mapu obarvit, aniž by výsledek byl zkreslující nebo zavádějící. Například politické mapy jsou často obarvené systémem všechno, nebo nic, takže není poznat, že kandidát v daném regionu vyhrál třeba o jediné procento. Barvy nemusí nutně spadat do předem připravených škatulek, při citlivém přístupu fungují dobře barevné gradienty.

A nezapomínejte, že vás nikdo nenutí v článku využít všechna dostupná data. Začněte v malém a přidávejte jen tehdy, když je to nevyhnutelně potřeba.

8

tipy pro vizualizaci dat

Projděte si data ze všech úhlů

Při analýze dat neexistuje žádná špatná perspektiva. Vyzkoušejte každý úhel pohledu, který vás napadne. Když píšete o zločinu, může se vám hodit graf meziročního vývoje násilných zločinů, další pohled nabízí procentuální změna nebo srovnání s jinými městy. Vyzkoušejte absolutní čísla, procenta, indexy.

Prohlédněte si data v různých měřítcích. Zkuste si umístit osu x tradičně na nulu, pak ji posuňte jinam. Změnilo se něco? Pokud mají data nepraktické rozložení, můžete je logaritmovat nebo odmocnit.

Každá taková změna vám pomáhá vidět data v novém světle. Jakmile vám přestanou říkat něco nového, máte hotovo.

Ne každá chyba je fatální

Když si data prohlédnete ze všech úhlů, určitě narazíte na čísla, která nehrají. Možná jim vůbec nerozumíte, možná se výrazně liší od zbytku souboru, možná jde o překlepy, možná neodpovídají trendům.

Pokud na takových datech hodláte postavit článek nebo je chcete publikovat, musíte se s anomáliemi nějak vypořádat. Výjimka v datech může být skvělý námět na zajímavý článek i obyčejná chyba. Zajímavá výzva pro zaběhnuté názory i pouhé nedorozumění.

Pokud data pochází od státu, chyby v nich bývají celkem běžně. Stejně tak je velmi jednoduché špatně pochopit nějaký úřednický výraz.

V první řadě zkuste zkontrolovat svou vlastní práci. Přečetli jste si dokumentaci, varovala vás před něčím? Objevuje se problém i v původní, nezpracované verzi dat? Pokud na vaší straně vypadá všechno v pořádku, je čas zvednout telefon. Chcete-li data použít, nějak se s chybou vypořádat musíte, tak proč ne hned.

Na druhou stranu není každá chyba zásadní. Například v záznamech o financování volební kampaně se běžně stává, že se mezi stovkami tisíc položek najde několik set neexistujících PSČ. Pokud všechny takové záznamy nepatří do jednoho regionu nebo jednomu kandidátovi, nemá smysl si s občasnou chybou dělat hlavu.

Základní otázka zní, jestli chyby v datech zásadně zkreslují dojem, který z nich získají vaši čtenáři.

Netrapte se nepřesnostmi

Netrapte se nepřesnostmi, dokud na nich opravdu nezáleží. Vaše experimentální průběžné vizualizace sice musí být v principu správně, ale nesejde na tom, jestli v nich všude používáte jednotné zaokrouhlení, jestli vám všechna procenta správně vychází přesně do stovky, nebo jestli vám mezi daty za dvacet let nechybí jeden dva roky. Drobné nepřesnosti jsou přirozenou součástí experimentu. Důležité je zachytit větší trendy a vědět, co ještě potřebujete před zveřejněním nasbírat a upřesnit.

Můžete dokonce zkusit vypustit popisky a měřítko, podobně jako na výše uvedených grafech, a nerušeně se zabývat jen celkovým tvarem dat.

Kdy se vizualizace nehodí

Efektivní vizualizace vyžaduje rozumně kvalitní, čistá, přesná a smysluplná data. Podobně jako se klasická novinařina opírá o kvalitní citace, fakta a popisy, i datová vizualizace je jen tak dobrá jako data, ze kterých vychází. Kdy je lepší použít jiné nástroje?

Když se váš příběh víc hodí pro text nebo multimédia. Některé příběhy se v číslech vypráví špatně. Jednoduchý graf dokreslující trendy je pěkná věc, stejně jako shrnující statistika. Ale pro bezprostřední, úderný popis některých problémů a jejich dopadu na reálný svět je nejlepší text.

Když máte málo dat. Jedno číslo samo o sobě nic neznamená. V reakci na citované statistiky bývá od editorů často slyšet otázka: „Ve srovnání s čím?“ Jde trend nahoru, nebo dolů? Jak vypadá normální stav?

Když v datech schází jasný pohyb. Při pohledu na data vykreslená například v Excelu občas zjistíte, že jsou plná šumu, hodně kolísají, chybí jim jasný trend. Co s tím? Začnete posouvat osy a měnit měřítko, aby křivka byla zajímavější? Ne! Nejspíš vám schází jednoznačná data, musíte se vrátit k analýze a najít lepší.

Když mapa není mapa. Pokud rozložení dat v prostoru nenese smysluplnou nebo zásadní informaci, jen odvádí pozornost od relevantnějších ukazatelů, například změny proměnných v čase nebo rozdílů mezi regiony, které spolu na mapě nesousedí.

A nezapomínejte na tabulky. Když máte několik málo čísel, která by ovšem mohla být pro čtenáře zajímavá, zkuste tabulku. Je srozumitelná a nevzbuzuje v čtenářích přehnaná očekávání nějakého grandiózního příběhu.

9.

dátová žurnalistika na slovensku

Na ďalších stranách prinášame krátku históriu dátovej žurnalistiky na Slovensku. Bude pravdepodobne neúplná a takmer určite subjektívna.

To najmä preto, že som asi ako prvý v slovenských médiách na konci roka 2012 začínal na SME.sk s niečím, čo sa už vtedy v iných krajinách volalo dátová žurnalistika. V prvých krokoch to znamenalo odsledovať, čo sa robí v Guardiane, ktorý bol vtedy priekopníkom. Neskôr to bolo hľadanie inšpirácií po Githuboch a Twitteroch a snaha o zmysluplnejšie dátové analýzy. Od začiatku roka 2014 slovenskú dátovú žurnalistiku pozorujem (a čiastočne spoluvytváram) už z prostredia mimo redakcie.

Matej Hruška

Začiatkom novembra 2015 sa v Prahe prvýkrát konal Editors Lab organizovaný GEN (Global editors network). Editors Lab je séria podujatí, kde sa zídu tímy novinárov, grafikov či developerov a vymýšľajú inovatívne spôsoby podávania informácií. Jednotlivé podujatia po svete organizujú rôzne redakcie spolu s ďalšími partnermi. V Prahe to boli Česká televízia a Google.

Témou bolo Covering the Refugees and Migrants Crisis with Data a zadanie by sa dalo zhrnúť takto: Navrhnite vizualizáciu, dátovú analýzu, kvíz, hru alebo čokoľvek, čo by pomohlo lepšie porozumieť cestám utečencov alebo ich ekonomickej a sociálnej situácii.

Víťazom pražského Editors Labu sa stal tím slovenskej RTVS a ich zaujímavý a užitočný projekt Štát na to má. Paradoxom pritom je, že RTVS nemá nijaký verejný online priestor, kde by sa podobným aktivitám (hľadaniu spôsobov, ako užitočne dáta spracovať) systematicky venovali, nijaký dátový tím. Celý úspech stojí na niekoľkých nadšencoch z redakcie. To by mohla byť aj prvá zjednodušená charakteristika stavu dátovej žurnalistiky na Slovensku v roku 2015 – sú tu skôr nadšenci pre dátovú žurnalistiku ako dátová žurnalistika samotná. Na ilustráciu – hádajte, čo nájdete na webe datovazurnalistika.sk? Nič.

Na slovenskú dátovú žurnalistiku narazíte v roku 2015 skôr na Facebooku cez rôzne voľnočasové alebo treťosektorové projekty, ako v slovenských médiách. Nie, že by v nich o niečom takom nepočuli. Niekoľko fanúšikov dát je roztrúsených po rôznych redakciách (okrem tímu RTVS v Prahe súťažil zo Slovenska aj tím SME), na väčšiu inštitucionálnu podporu a tomu zodpovedajúcu kvalitu výstupov však títo fanúšikovia zatiaľ čakajú. V skratke: Ako sci-fi u nás vyzerá nie New York Times, Vox či Guardian, ale už aj české Sami z dat.

Nie je táto melancholická lamentácia iba obhajobou vlastného záujmu? Potrebujú noviny svojich dátových novinárov a novinárky?

Honza Boček (z vyššie spomínaných samizdat) píše v českej verzii tejto príručky: „Datová žurnalistika je něco jiného. Je to ambice podívat se na mediální témata hlouběji, ptát se na neobvyklé otázky, hledat celkový obraz, vysvětlovat složité a komplikovat jednoduché problémy, většinou pomocí dat.“ Táto ambícia by mala byť vlastná aj dnešným médiám. Samozrejme, neplatí to úplne pre všetky a nie všetky si napĺňanie tejto ambície môžu dovoliť. Ako píše Honza, ich tím po dvoch rokoch skončil v súkromných novinách („Naše články vznikaly příliš dlouho, než aby se vydavateli vyplatily“) a presunul sa do verejnoprávneho Českého rozhlasu.

Why so serious? Mimovládky zasahujú

Slovenskú systematickú, analytickú a vizuálnu prácu s dátami je najlepšie predstaviť cez aktivity mimovládok a tretieho sektora. V neúplnom prehľade je dobré spomenúť:

Aliancia Fair-play sa okrem vlastnej investigatívy stará o Datanest a (spolu)organizuje hackathony. Portál znasichdani.sk získal v najväčšej celoeurópskej dátovej súťaži Open Data Challenge 2011 ocenenie za najlepšiu európsku open-data aplikáciu pracujúcu s verejnými informáciami.

Transparency International Slovensko stojí okrem iných projektov aj za portálom otvorenesudy.sk službou tender.sme.sk (taktiež medzinárodne ocenenou). TIS a AFP spolu vymysleli otvorenezmluvy.sk.

INEKO zbiera a vyhodnocuje dáta o školách, nemocniciach, samosprávach. Vytvorili „hru“ na ministra financií, ktorá dostala aj anglickú verziu a možnosť upraviť ju aj pre iné štáty.

V kohovolit.eu analyzujú hlasovania poslancov, pripravujú volebné kalkulačky a volebné mapy, porovnávajú aktivity europoslancov a účasť ministrov na zasadnutiach Rady EÚ.

V tomto výpočte ide často o príklady užitočných nástrojov a nie o dátovú žurnalistiku. Práve tieto nástroje tvoria kostru užitočnej práce s dátami na Slovensku, ktorá ide nad rámec vyrábania grafov do novín. Dávajú možnosť pri témach, ktoré sú pre krajinu dôležité (zdravotníctvo, školstvo, súdy, obstarávanie a pod.), odpovedať aj na nezvyčajné a komplikované otázky. Slovenské mimovládky sa stávajú v niektorých témach viac „data-driven“, čo vplýva aj na pokrývanie a vizuálnu prezentáciu týchto tém v médiách.

Za posledné mesiace by sme našli viac príkladov analytickej spolupráce tretieho sektora a médií pri odhaľovaní rôznych káuz – napríklad v zdravotníctve a pri téme schránkových firiem. Možno práve to bude cesta pre slovenské médiá, ako bez výraznejšieho rastu nákladov naplniť vyššie opísanú ambíciu.

Čo ďalej

Dátovú žurnalistiku nemusia tvoriť iba vážne témy. Samozrejme, najlepšie by bolo, ak by sme mali poctivo pokryté aspoň tie. Analytika a práca s dátami dobre sedí aj témam z kultúry, športu, technológií či životného prostredia, kde sa dá ísť za hranice tradičných žánrov. Tu sa ponúka spomenúť zahraničné príklady (hoci sú v každom podobnom texte rovnaké): Vox (a ich tagline Explain the news), 538 alebo Quartz. Z iného spektra na hraniciach dátového bulváru napríklad Ampp3d patriaci pod britský Daily Mirror, ktorý sa na Twitteri opisuje: The numbers tell the REAL story.

10.

mladý data-žurnalistický talent, ondrej proksa

Nestačí mať len dáta, treba mať aj nápad

Ondrej Proksa sa začiatkom roka 2014 prihlásil do prvého ročníka Ceny Google pre mladé talenty dátovej žurnalistiky, ktorá sa udeľuje v rámci Novinárskej ceny Nadácie otvorenej spoločnosti. A to hneď s viacerými námetmi. Nečakal, že jeden z nich bude naozaj úspešný. Porotu najviac zaujal jeho nápad spracovať dáta z volieb do prehľadnej podoby, nájsť v nich nové prepojenia a vzťahy, ktoré dovtedy verejnosti unikali. Svoj projekt neskôr realizoval v spolupráci s webovým vydaním denníka Pravda. Vďaka tejto spolupráci k voľbám pribudli i dáta z referend.

Čo vás inšpirovalo pozrieť sa práve (na dáta z volieb?

Volebné dianie na Slovensku ma dlhodobo zaujíma. Rád sledujem predvolebné kampane, priebeh a výsledky volieb. Počas minulého „super“ volebného roku som sa rozhodol, že spolu s kolegom spracujem výsledky volieb zo ŠÚ SR (Štatistického úradu Slovenskej republiky). Na Slovensku nám chýbal portál, ktorý by sumarizoval všetky výsledky volieb. Okrem toho mi chýbal nástroj, pomocou ktorého by ste si mohli vyhľadať nejakú obec a pozrieť všetky výsledky z rôznych volieb. Takto by sa napríklad mohlo odhaliť kupovanie hlasov alebo sledovať, ako vplýva starosta (alebo jeho stranícka príslušnosť) na výsledky v prezidentských voľbách.

Počas hlbšej analýzy danej problematiky sme narazili aj na historické volebné dáta z roku 1920, ktoré vlastní Sociologický ústav SAV. Tieto volebné dáta nie sú nikde prezentované a nie je jednoduché ich spracovať a vizualizovať. Práve vďaka tomuto zámeru náš nápad v súťaži zvíťazil.

Na aké problémy ste pri realizácii vášho projektu narazili? (Predovšetkým z pohľadu získavania a spracúvania dát.)

Počas realizácie projektu som narazil na viacero problémov: preklepy, nekonzistentnosť alebo zlé názvy obcí či politických strán.

Najväčším problémom bola nekonzistentnosť dát. Spracúval som výsledky od roku 1990 do roku 2014. Počas tohto obdobia sa menila aj organizačná štruktúra samospráv. Vznikalo viacero nových obcí, menil sa počet mestských častí v Bratislave a Košiciach, prípadne niektoré obce sa spojili alebo zanikli. Za zmienku stojí to, že som ručne musel opravovať prepojenia, aby som zachoval hierarchiu a výsledky pre Bratislavu boli kompletné.

Ďalším problémom bola nejednoznačnosť v dátach – viacero obcí na Slovensku má rovnaký názov. Pri starších výsledkoch nebolo k obci žiadne iné označenie (ani kód, ani okres, resp. kraj). Výsledky som musel prepájať ručne, prípadne vymýšľať metriky, ktoré obce spárujú správne (napr. podľa počtu voličov ku počtu obyvateľov).

Pri volebných dátach som sa trápil aj s politickými stranami. Strany často menia svoje názvy, prípadne menia úplne svoje identity (napr. predaj strany). V takýchto prípadoch som musel o pomoc žiadať priamo novinárov, aby našli v archívoch zmienky o histórii strán.

Čo by podľa vás mohlo pomôcť tieto problémy vyriešiť alebo aspoň zmierniť?

Historické dáta je často náročné optimalizovať a začisťovať. V súčasnosti ŠÚ SR už pridáva k obciam kódy a tiež identifikácie okresu a kraju. Vďaka tomu možno identifikovať jedinečnú obec na Slovensku. Stále však platí, že ak údaje spracúvajú ľudia, riziko chýb a preklepov je možné.

Ako víťaz Ceny Google pre mladé talenty dátovej žurnalistiky ste mohli svoj projekt zrealizovať pod vedením redakcie Pravdy. Čo vám spolupráca s profesionálnymi novinármi priniesla?

Verejné dáta som poznal, rád som sa so spracovanými dátami hral, ale nikdy som sa nad tým nezamýšľal z pohľadu žurnalistiky. Práve po tomto víťazstve v súťaži sa mi otvorili nové obzory. Spolupráca s Pravdou ma posunula dopredu. Naučil som sa viacej rozumieť dátovej žurnalistike z novinárskej perspektívy. Pochopil som, že nestačí mať len dáta, ale treba vymyslieť, ako ich správne vizualizovať. Zistil som, že ak máme nevyčistené dáta, napríklad archívne dáta z výsledkov volieb, práca s nimi nie je taká jednoznačná, ako sa na prvý pohľad môže zdať.

Podarilo sa nám spolu vytvoriť nový projekt na zobrazenie, analýzu a spracovanie volebných výsledkov. Okrem pracovnej roviny vznikli vďaka tejto súťaži aj priateľstvá. V Pravde som najviac komunikoval s Mariánom Nitonom, s ktorým som si neskôr aj potykal.

Spolupráca s Pravdou mi priniesla ďalšie možnosti spolupráce aj s inými inštitúciami – napr. Transparency International alebo Alianciou Fair-Play.

Našli ste v analyzovaných dátach niečo, čo vás prekvapilo?

Všetky verejné dáta sú zaujímavé a pre novinárov cenné. Často sa v nich nachádzajú zaujímavé analytické odhalenia alebo prepojenia. Pri volebných výsledkoch som vymyslel tzv. časovú os obce, ktorá zobrazí na jednom mieste všetky výsledky volieb. Na časovej osi možno sledovať, ako sa menia starostovia, ako sa mení vládna strana, prípadne aká strana vyhrala prezidentské voľby v danej obci. Táto sumarizácia môže priniesť rôzne zaujímavé odhalenia.

Mne sa podarilo nájsť starostov, ktorí boli zvolení 5–6krát. Našiel som aj obce, ktoré si každé volebné obdobie zvolili nového starostu.

Podporila škola váš zámer odhaľovať súvislosti vo verejne dostupných dátach?

Pred prvým ročníkom, do ktorého som sa zapojil, sme mali prednášku o dátovej žurnalistike na škole. Dekanka FIIT prof. Mária Bieliková nás nabádala, aby sme sa do súťaže prihlásili. Môžem povedať, že vďaka jej iniciatíve som sa o súťaži dozvedel a mohol sa zapojiť. Okrem iného, po propagácii na škole sa do oboch ročníkov prihlásilo viacero študentov FIIT so svojimi nápadmi.

Aké dáta vás lákajú do budúcna, prípadne o aké dáta sa zaujímate teraz?

Minulý rok som sa do súťaže prihlásil s nápadom v oblasti verejného obstarávania. Počas skúmania a spracúvania dát som objavil možné prepojenia medzi víťazmi verejného obstarávania a komisiou, ktorá ponuky vyhodnocovala. Tento nápad bol úspešný, a preto som sa rozhodol venovať mu ďalej. Hľadám partnera – médium alebo mimovládnu organizáciu – s ktorým budeme projekt realizovať.

Okrem dát z verejného obstarávania by som rád analyzoval dáta samospráv. Každá samospráva zverejňuje svoje hospodárenie, rozpočty a tiež zmluvy. Problém v týchto dátach je v tom, že každá samospráva zobrazuje údaje v inej forme, iným spôsobom a v inom rozsahu. Bolo by zaujímavé vedieť automatizovane porovnávať rozpočty, zmluvy, prípadne iné vzťahy v obciach a mestách. Jedným z cieľov by mohlo byť hľadanie a sledovanie toku financií do obcí, kde sú rôzne stranícke príslušnosti (vládna strana, resp. strana, ktorá je v opozícii) a pod. Možno som práve načrtol nápad do ďalšieho ročníka súťaže. :)

- Ondrej Proksa je absolventom Fakulty informatiky a Informačných technológií Slovenskej technickej univerzity v Bratislave. Dnes sa ako freelancer venuje oblasti vývoja softvéru a dát. Stal sa víťazom Ceny Google pre mladé talenty dátovej žurnalistiky za roky 2013 i 2014 udeľovanej v rámci Novinárskej ceny.

- Cena Google pre mladé talenty dátovej žurnalistiky je určená študentom a mladým ľuďom do 30 rokov, ktorí sa prihlásia so zaujímavým nápadom, ako využiť a zanalyzovať verejné dáta. Víťaz v tejto kategórii získava finančnú odmenu a možnosť realizovať svoj projekt v jednom z profesionálnych médií.

Odporúčané nástroje

- Open Refine, open-source aplikácia na čistenie dát.
- Google Fusion Tables, dynamické tabuľky, ktoré umožňujú kombinovať a používať dáta z rôznych zdrojov a aplikácií. Zvládajú prácu s dátami do 100 megabajtov a základné grafy a mapové vizualizácie.
- Tableau Public, softvér na analýzu dát a ich interaktívnu vizualizáciu. V základnej verzii je na spracovanie datasetov do 100-tisíc riadkov, súčasťou webu sú aj tutorial videá a návody.
- Datawrapper, open-source nástroj na tvorbu jednoduchých grafov.
- Datamatic, český start-up, ktorý vyvinul nástroj na tvorbu vizualizácií.

Odporúčané zdroje

- The Data Journalism Handbook, pôvodná verzia príručky dátovej žurnalistiky (220 strán), na ktorej je táto publikácia postavená. Webová verzia je k dispozícii zadarmo.
- DataDrivenJournalism.net, web mapujúci trendy v dátovej žurnalistike. Jeho súčasťou je aj komunitný mailinglist združujúcu odborníkov a nadšencov z celého sveta.
- Doing Journalism with Data: First Steps, Skills and Tools, bezplatný on-line kurz základov žurnalistiky, vedený špičkami odboru.
- Making Sense of Data, úvod do dátovej gramotnosti z dielne Googlu. Zadarmo a on-line.
- Záznam workshopu dátovej žurnalistiky vedený Karlom Minaříkom a Josefom Šlerkom. Záznam 1, 2, 3, 4.
- School of Data, projekt OKFN, ktorý sprostredkúva návody na základnú aj pokročilejšiu prácu s dátami.

Autoři The Data Journalism Handbook

- Gregor Aisch (Open Knowledge Foundation)
- Brigitte Alfter (Journalismfund.eu)
- David Anderton (Freelance Journalist)
- James Ball (The Guardian)
- Caelainn Barr (Citywire)
- Mariana Berruezo (Hacks/Hackers Buenos Aires)
- Michael Blastland (Freelance Journalist)
- Mariano Blejman (Hacks/Hackers Buenos Aires)
- John Bones (Verdens Gang)
- Marianne Bouchart (Bloomberg News)
- Liliana Bounegru (European Journalism Centre)
- Brian Boyer (Chicago Tribune)
- Paul Bradshaw (Birmingham City University)
- Wendy Carlisle (Australian Broadcasting Corporation)
- Lucy Chambers (Open Knowledge Foundation)
- Sarah Cohen (Duke University)
- Alastair Dant (The Guardian)
- Helen Darbishire (Access Info Europe)
- Chase Davis (Center for Investigative Reporting)
- Steve Doig (Walter Cronkite School of Journalism of Arizona State University)
- Lisa Evans (The Guardian)
- Tom Fries (Bertelsmann Stiftung)
- Duncan Geere (Wired UK)
- Jack Gillum (Associated Press)
- Jonathan Gray (Open Knowledge Foundation)
- Alex Howard (O'Reilly Media)
- Bella Hurrell (BBC)
- Nicolas Kayser-Bril (Journalism++)
- John Keefe (WNYC)
- Scott Klein (ProPublica)
- Alexandre Léchenet (Le Monde)
- Mark Lee Hunter (INSEAD)
- Andrew Leimdorfer (BBC)
- Friedrich Lindenberg (Open Knowledge Foundation)
- Mike Linksvayer (Creative Commons)
- Mirko Lorenz (Deutsche Welle)
- Esa Mäkinen (Helsingin Sanomat)
- Pedro Markun (Transparência Hacker)
- Isao Matsunami (Tokyo Shimbun)
- Lorenz Matzat (OpenDataCity)
- Geoff McGhee (Stanford University)
- Philip Meyer (Professor Emeritus, University of North Carolina at Chapel Hill)
- Claire Miller (WalesOnline)
- Cynthia O'Murchu (Financial Times)
- Oluseun Onigbinde (BudgIT)
- Djordje Padejski (Knight Journalism Fellow, Stanford University)
- Jane Park (Creative Commons)
- Angélica Peralta Ramos (La Nacion, Argentina)
- Cheryl Phillips (The Seattle Times)
- Aron Pilhofer (New York Times)
- Lulu Pinney (Freelance Infographic Designer)
- Paul Radu (Organised Crime and Corruption Reporting Project)
- Simon Rogers (The Guardian)
- Martin Rosenbaum (BBC)
- Amanda Rossi (Friends of Januária)
- Martin Sarsale (Hacks/Hackers Buenos Aires)
- Fabrizio Scrollini (London School of Economics and Political Science)
- Sarah Slobin (Wall Street Journal)
- Sergio Sorin (Hacks/Hackers Buenos Aires)
- Jonathan Stray (The Overview Project)
- Brian Suda (optional.is)
- Chris Taggart (OpenCorporates)
- Jer Thorp (The New York Times R&D Group)
- Andy Tow, Hack (Hackers Buenos Aires)
- Luk N. Van Wassenhove (INSEAD)
- Sascha Venohr (Zeit Online)
- Jerry Vermanen (NU.nl)
- César Viana (University of Goiás)
- Farida Vis (University of Leicester)
- Pete Warden (Independent Data Analyst and Developer)
- Chrys Wu (Hacks/Hackers)

2011	01	**Nadobúda účinnosť zákon o zverejňovaní zmlúv na webe** **Vzniká vládny Centrálny register zmlúv (CRZ)**
	02	
	03	**AFP Aliancia Fair-Play (AFP) rozbieha projekt znasichdani.sk**
	04	
	05	
	06	
	07	
	08	
	09	
	10	**Transparency International Slovensko a AFP spúšťajú otvorenezmluvy.sk**
	11	Zrození konceptu The Data Journalism Handbook (Mozilla Festival)
	12	**Vzniká Cena Google za inovatívnu online žurnalistiku, ktorá sa udeľuje pravidelne v rámci Novinárskej ceny**
2012	01	**Vzniká Register účtovných uzávierok**
	02	
	03	**Vzniká vládny portál pre otvorené dáta data.gov.sk**
	04	Představení hotové publikace (Perugia)
	05	
	06	
	07	Publikace v distribuci (O'Reilly, Amazon)
	08	
	09	Vydání ruské mutace
	10	
	11	Mapujeme zájem o české vydání
	12	

2013	01	Získáváme finanční podporu od Velvyslanectví USA v Praze
	02	
	03	Pořádáme workshop o datové žurnalistice **Vzniká projekt Transparency International SK otvorenesudy.sk**
	04	Vydání španělské mutace Začínáme pracovat na českém vydání příručky
	05	Vyčlenění financí z grantu CEE Trustu
	06	Překlad, editace, vznik českých autorských textů
	07	Design, sazba, korektury
	08	Vychází francouzská mutace Guide du datajournalisme Vydáváme tištěnou příručku a e-knihu v češtině
	09	
	10	
	11	Částečný překlad do gruzínštiny
	12	1. ročník Ceny Googlu pro mladé talenty datové žurnalistky **1. ročník Ceny Google pre mladé talenty dátovej žurnalistiky**
2014	01	
	02	
	03	
	04	1. vítěz ceny pro datažurnalistické talenty Petr Zvirinský s projektem Insolvenční rejstřík
	05	**1. víťazom Ceny Google pre mladé talenty dátovej žurnalistiky sa stáva Ondrej Proksa s projektom volby.pravda.sk**
	06	
	07	
	08	
	09	Vydání ruské mutace
	10	
	11	
	12	2. ročník Ceny Googlu pro mladé talenty datové žurnalistiky **2. ročník Ceny Google pre mladé talenty dátovej žurnalistiky**

2015 01
02
03
04
05
06
07
08
09
10
11
12
2016 01
02
03
04
05
06
07
08
09
10
11
12

1. vítězka ceny pro datažurnalistické talenty Petra Paříková

AFP spúšťa projekt pozemky.fair-play.sk

2. víťazom Ceny Google pre mladé talenty dátovej žurnalistiky sa stáva Ondrej Proksa

Vzniká nová verzia katastra nehnuteľností – Mapka

Vzniká aplikácia na prehliadanie údajov katastra nehnuteľností „CICA“

2. upravené vydání české verze
Příručky datové žurnalistiky

Nadácia otvorenej spoločnosti s partnermi v Česku, Maďarsku a Rumunsku spúšťa projekt Dáta menia žurnalistiku

Prvé vydanie slovenskej verzie Príručky dátovej žurnalistiky

2017 01
02
03
04
05
06
07
08
09
10
11
12
2018 01
02
03
04
05
06
07
08
09
10
11
12

ISBN 978-80-87725-35-1

překlad a editace
tomáš znamenáček

autorské texty
jan boček
jan cibulka
petr kočí
michaela rybičková
adam valček
matej hruška
ľubica stanek

grafická úprava
ex lovers

Základem tohoto textu byla publikace The Data Journalism Handbook, která byla přeložena, zkrácena a doplněna autorskými texty z českého prostředí.