

Příručka datové žurnalistiky

20. 6. 2013

Obsah

Předmluva	3
Úvod	4
Proč je datová žurnalistika důležitá	4
Informační filtr	4
Budoucnost žurnalistiky	5
Nový nástroj	5
Odpověď na datové PR	5
Nezávislá interpretace oficiálních dat	6
Data jsou náš život	6
Časová úspora	6
Nezbytná součást výbavy	7
Adaptace na změny v našem informačním prostředí	7
Datový vesmír	7
Otevřená data pro zpětnou kontrolu	8
Datová gramotnost ve třech krátkých krocích	8
Odkud se data vzala?	8
Fantastický růst HDP	8
Věčně rostoucí křivka zločinu	9
Co můžete udělat	9

Co přesně data říkají?	9
Noční práce zdvojnásobuje riziko roztroušené sklerózy	9
Mezi každými 15 Evropany je průměrně jeden negramotný	10
Co můžete udělat	10
Jak spolehlivá jsou vaše data?	10
Problematická velikost vzorku	10
Pít čaje snižuje riziko infarktu	10
Co můžete udělat	11
Tipy pro práci s čísly	11
Hledejte příběh	11
Vnímejte kontext	12
Užívejte si	12
Skepse ano, cynismus ne	12
Nejdřív data, potom závěry	13
Nejistota není sprosté slovo	13
Psaní je příběh	13
Extrakce dat z webu	13
Co jsou strojově čitelná data	14
K čemu je scrapování	14
Scrapovací nástroje	15
Technické principy scrapování	15
Tipy pro práci s internetovými zdroji	16
WHOIS	16
Google Site Search	16
Google Cache	17
Webový archiv	17
Hledání obrázků	17
Google Trends	17

Vizualizace jako tažný kůň datové žurnalistiky	18
Různé grafy, různé příběhy	18
Tipy pro vizualizaci dat	23
Projděte si data ze všech úhlů	23
Ne každá chyba je fatální	24
Netrapte se nepřesnostmi	24
Kdy se vizualizace nehodí	24
Datová žurnalistika v Česku	25
Data v novinách: popis jednoho zápasu	25
Když novináři programují	26
Pasti a pastičky	26
Jako zrušení cenzury	27
Happy end?	27
Přístup k informacím z pohledu uživatele	28
Infozákon a paní na ústředně	28
Zajímejte se	29
Když to jinak nejde	29
Bitva paragrafů	30
Doporučené nástroje a zdroje	31
Nástroje	31
Zdroje	31
Tiráž	32

Předmluva

Datová žurnalistika, trend světových newsroomů, si do tuzemských redakcí teprve hledá cestu. Česká verze [The Data Journalism Handbook](#) tento rychle se rozvíjející fenomén přibližuje českému publiku a nabízí různorodý pohled na nové způsoby práce s daty a možnosti jejich využití v novinářské praxi. Z publikace, která je společným dílem desítek novinářů i datových nadšenců z celého světa, jsme vybrali nejpřínosnější pasáže a opatřili je českým kontextem. Cílem příručky není nabídnout kompletní návod, ale

spíše úvod do problematiky plný inspirativních příkladů a praktických tipů. Věříme, že bude pro české novináře tím pravým průvodcem na dobrodružnou výpravu krajinou velkých dat.

—Michaela Rybičková, Fond Otakara Motejla

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.

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 nástroje, které tradiční žurnalistika postrádá: nástroje pro hledání v digitálních zdrojích, analýzu dat a jejich vizualizaci. Nechce tradiční novinářinu 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 naprostě 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

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 více 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 rozptýlu 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 nespolehat na jejich oficiální výklad.

—Isao Matsunami, *Chunichi/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 slepuou uličku, občas žádné jiné ani nejsou. Ostatně – kdyby stačilo jen zmáčknout ten správný čuďák, 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, hacker, *Transparência*

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 nejhodnotnějším 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 Slobinová, *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 novinářinu, na které záleží, tohle všechno jsou plusy.

—Cheryl Phillipsová, *Seattle Times*

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 pář 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 dají i tak diskutované údaje, 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 nejvíce jednoduchých případů nevyžadujících vyšetřování. Například kouření marihuany. Tím se vysvětluje, proč ve Francii za posledních 15 let statisticky vzato čtyřikrát přibylo trestních 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 hroznivě. 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 rozdelená 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 bourá často, čímž tlačí aritmetický průměr nehodovosti mnohem výš, než by běžný řidič ze své zkušenosti čekal. Totéž platí o rozdelení příjmů: většina lidí má podprůměrný plat (a nadprůměrný počet končetin).

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 rádech 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 dostáhli 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 více, 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 korelace mezi čajem a sportováním.

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 utráví 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 prozměnu 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 milionu Euro 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 více než za rozvoz jídel pro seniory“) nebo vnějším srovnáním („dali loni za sponky dvakrát více, 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 více, než dělá republikový průměr.“

Také můžete data rozdělit na kategorie („úřady řízené stranou X utratí za sponky o polovinu více, než úřady obsazené stranou Y“), případně zdůraznit souvislosti: „Úřady řízené politiky, kteří dostali dary od výrobců kancelářských potřeb, utrácejí za kancelářské sponky více, 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 nepřistupně, ale když se jimi necháte zastrašit, nikam se nedostanete. Nebojte se s nimi pohrat, 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 více než příslušené 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 desítky 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 obrazek. Nesnažte se nutně působit jako autorita, která čtenářům naservíruje až hotovou odpověď.

Extrakce dat z webu

Ideální je, když se vám data podaří na webu najít v nějakém přímo zpracovatelném formátu, například jako excelovskou tabulkou 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 ted' 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 excelovské 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 tiskárna, 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 zkopirovat 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 postupovat 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 [DoS 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 webových 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 webovým 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 Documents 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 databázi, 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 mívala 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 *ScrapWiki*.

Tipy pro práci s internetovými zdroji

WHOIS

WHOIS je stručně řečeno registr vlastníků domén, IP adres a dalších internetových objektů. Existuje řada nástrojů, které s tímto registrem umí pracovat; jeden takový nabízí například český server [Lupa](#). V poslední době vlastníci domén často používají takzvanou soukromou registraci, při které se v registru neobjeví jejich podrobné údaje. V mnoha zbývajících případech ale můžete podle názvu domény v systému WHOIS zjistit jméno jejího vlastníka, jeho adresu, e-mail i telefonní číslo. Také můžete jako dotaz zadat IP adresu, podle které zjistíte informace o jednotlivci nebo organizaci, kterým počítač s touto adresou patří. To se výborně hodí například když se snažíte zjistit identitu uživatele webové služby, protože většina serverů si IP adresy svých návštěvníků zaznamenává a pamatuje.

Google Site Search

Při prohledávání obsahu konkrétní domény je nepostradatelným nástrojem vyhledávání Google a jeho klíčové slovo `site:`. Když ke svému dotazu přidáte řetězec `site:domena.cz`, Google vrátí pouze výsledky ze zadané domény. Dokonce můžete výsledky zúžit na konkrétní podadresy, například `site:domena.cz/stranky/`. Tento trik je zvlášť praktický při hledání informací, které vlastník domény sice zveřejnil, ale nehrne se do jejich propagace. Stačí trefit klíčová slova a můžete se dobrat velmi šťavnatých materiálů.

Google Cache

Kontroverzní stránky může jejich autor bez upozornění stáhnout nebo změnit. Pokud se potřebujete dostat k původnímu znění, můžete jako první instanci zkoušet kešovanou verzi, jak si ji při posledním indexování zapamatoval Google. Vyhledávač si vždy pamatuje pouze poslední verzi, takže musíte jednat rychle, než se přepíše tou aktuální. Zadejte do Googlu jako vyhledávací dotaz URL stránky a když se vám objeví ve výsledcích, najděte si u ní odkaz na kešovanou verzi (anglicky *cached*, česká verze používá označení *archiv*). Pokud uspějete, udělejte si snímek obrazovky nebo zkopírujte relevantní část obsahu; keš může být každým okamžikem aktualizována na současnou verzi stránky.

Webový archiv

Změny konkrétního serveru nebo stránky za delší časové období, řekněme měsíce nebo roky, si můžete prohlédnout pomocí služby [Wayback Machine](#), která pravidelně snímá velkou část webu. Stačí zadat adresu stránky, která vás zajímá, a pokud je v archivu, zobrazí se vám kalendář s vyznačenými historickými snímky. Po kliknutí na konkrétní den vám archiv ukáže obsah stránky zhruba v tehdejší podobě. Často chybí formátování nebo obrázky, ale základní představu o obsahu si uděláte snadno.

Hledání obrázků

Občas byste chtěli vědět, odkud pochází nějaký obrázek, ale klasické vyhledávače jako Google vám bez uvedení zdroje neporadí. Služba [TinEye](#) se specializuje právě na takové „hledání naruby“, kde zadáte obrázek a dostanete seznam webových stránek, na kterých se vyskytují podobné. Obrázky se srovnávají pomocí chytrých algoritmů, které zvládnou i menší rozdíly v ořezu, zkreslení nebo kompresi. Výborně se služba hodí v případech, kdy máte podezření, že už jste obrázek vydávaný za novinku nebo originál už někde viděli.

Google Trends

Jasný obrázek o tom, co lidé hledají na webu, si můžete udělat prostřednictvím služby [Google Trends](#), kde Google zveřejňuje své statistiky vyhledávaných frází. Můžete zadat jednu konkrétní frázi („[Fukushima](#)“) nebo více frází oddělených čárkou („[Karel Schwarzenberg](#), [Miloš Zeman](#)“) pro vzájemné srovnání. Výběr dat se dá zúžit podle různých kritérií, například zeměpisně nebo časově. Škoda jen chybějících absolutních čísel – graf ukazuje pouze relativní „zájem“ o danou frázi v procentech, která nemusí mít jasnou výpovídající hodnotu.

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 téma 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, zvláště 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.

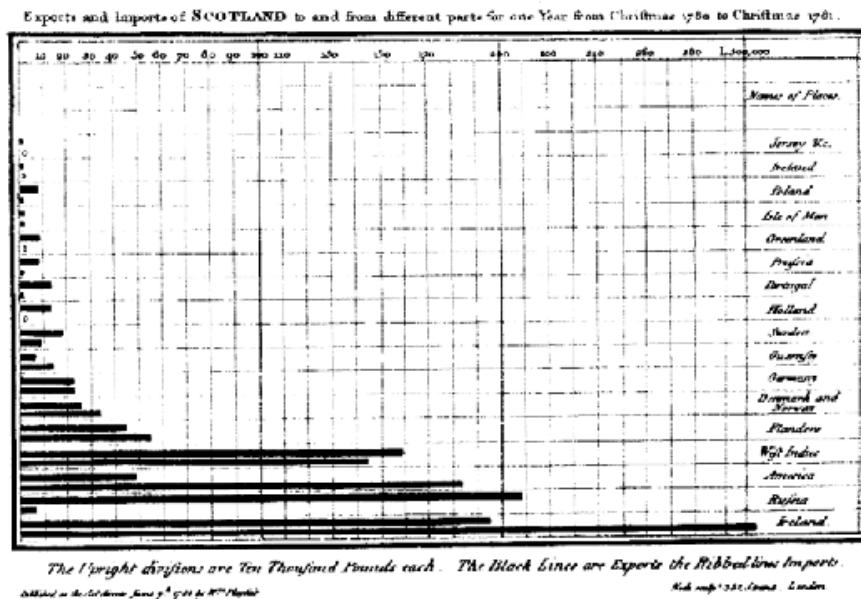
Různé grafy, různé příběhy

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 plachý papír dnes považujeme za médium druhé kategorie, ale faktum je, že za stovky let psání 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 nejznámnějších 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 sloupový graf, kterým nově a přehledně ilustroval skotský import a export.

Následovala kniha *Statistical Breviary*, kterou Playfair v roce 1801 popularizoval dnes také 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 nakresil 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.

V průběhu let se nový obor osměloval 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



Obrázek 1: Jeden z prvních sloupcových grafů; William Playfair

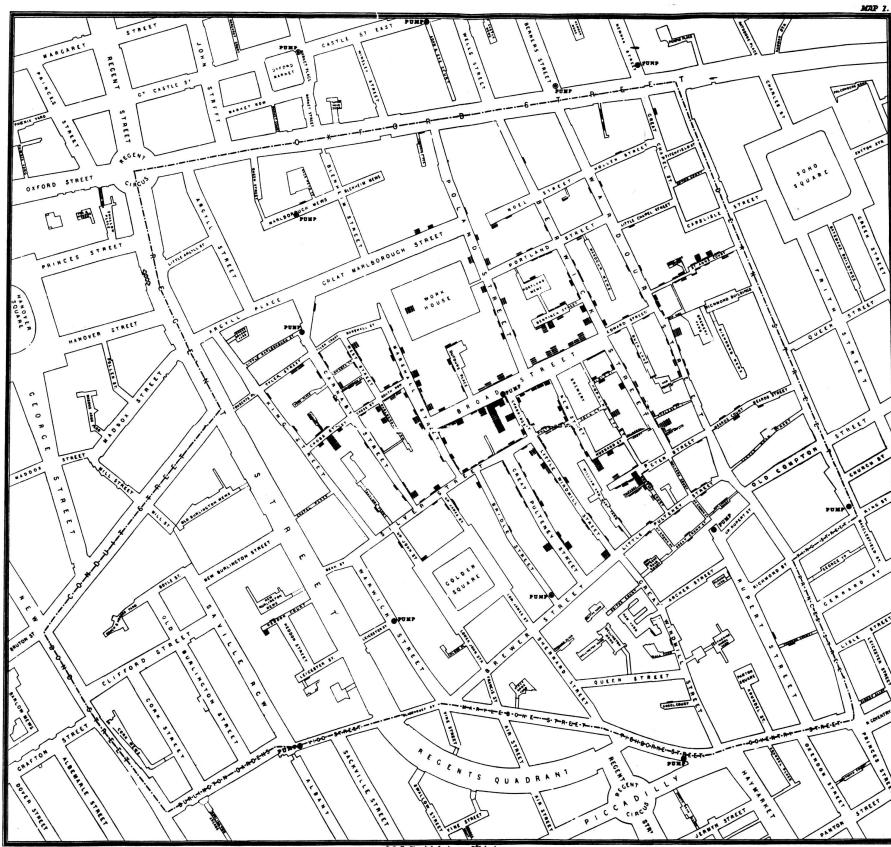
dalších ukazatelů vázaných na zeměpisnou oblast. Nápad je to v principu velmi jednoduchý, ale pokud nemá výsledná mapa zkreslovat a má být srozumitelná pro čtenáře, vyžaduje jisté úsilí.

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.

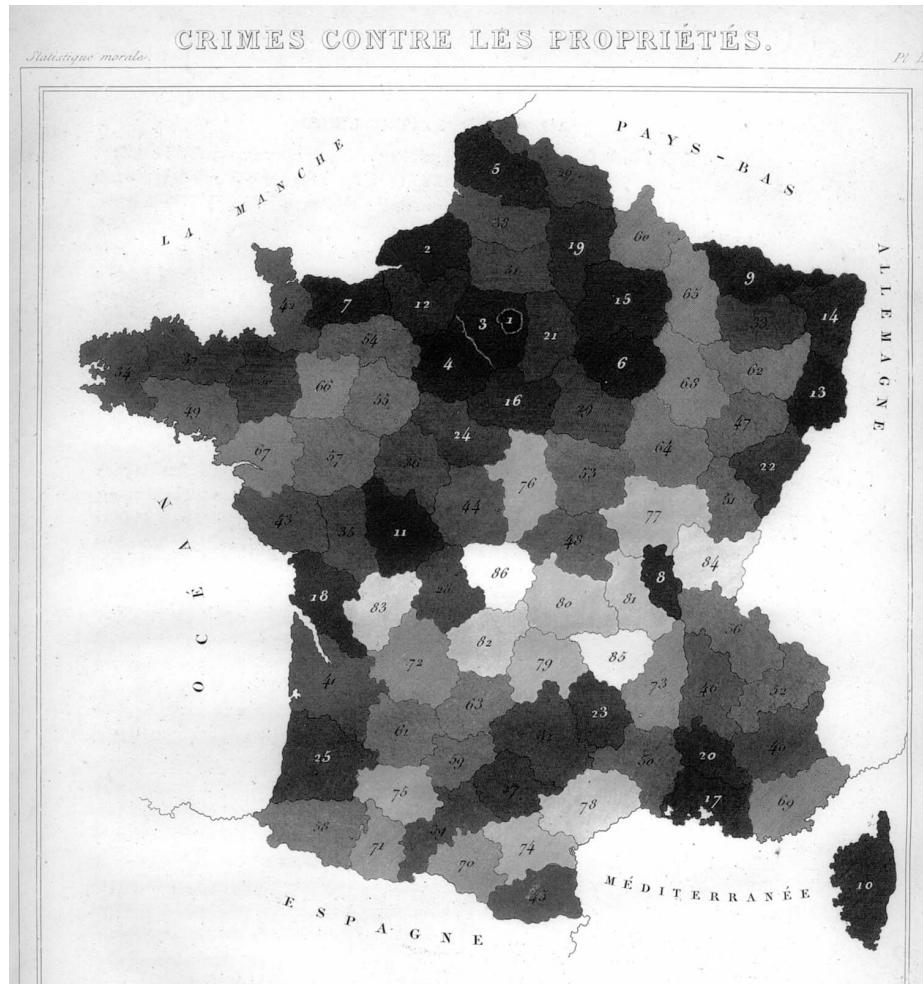
Mezi ty nejzákladnější typy grafů patří čárové a sloupcové grafy. Používají se v podobných případech, ale jsou mezi nimi i podstatné významové rozdíly. Podívejme se například na měsíční statistiku firemních příjmů za jeden rok. Při popisu sloupcovým grafem dostaneme 12 sloupečků, z nichž každý ukazuje 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 táz čísla, ale z průběhu čáry bychom mohli získat dojem, že někdy v polovině ledna firma vydělala \$110. Což není pravda. Pro nespojitá data se více hodí sloupcový graf; čárový graf je ideální pro data spojitá, například průběh teplot.

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 spojity – jednotlivé body grafu nepopisují součet



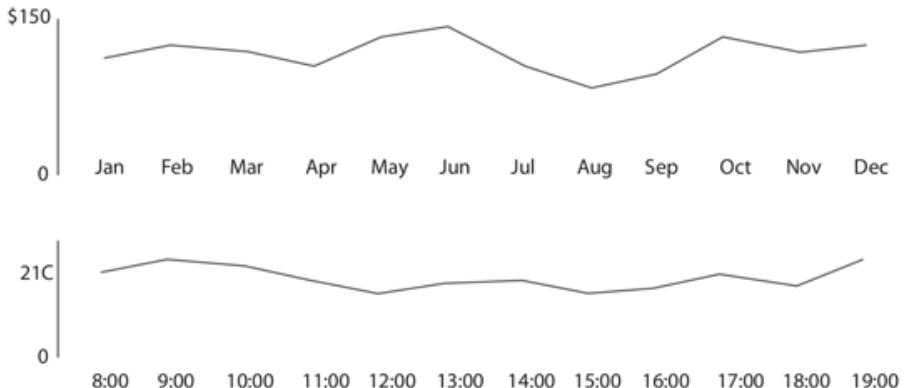
Obrázek 2: Cholerová mapa Londýna; John Snow



Obrázek 3: Kriminalita ve Francii; André-Michel Guerry



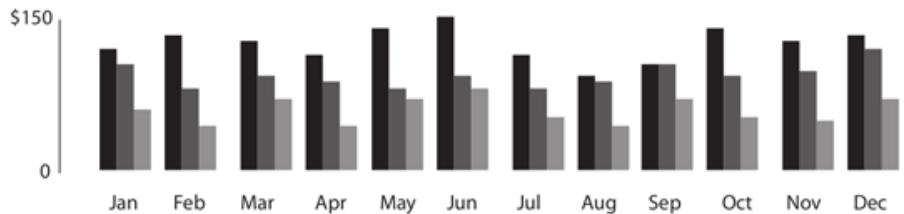
Obrázek 4: Jednoduchý sloupkový graf, ideální pro reprezentaci nespojitých dat



Obrázek 5: Jednoduchý čárový graf, ideální pro reprezentaci spojitého dat

nějakých čísel, nýbrž konkrétní hodnotu v daném čase nebo její odhad mezi dvěma měřeními.

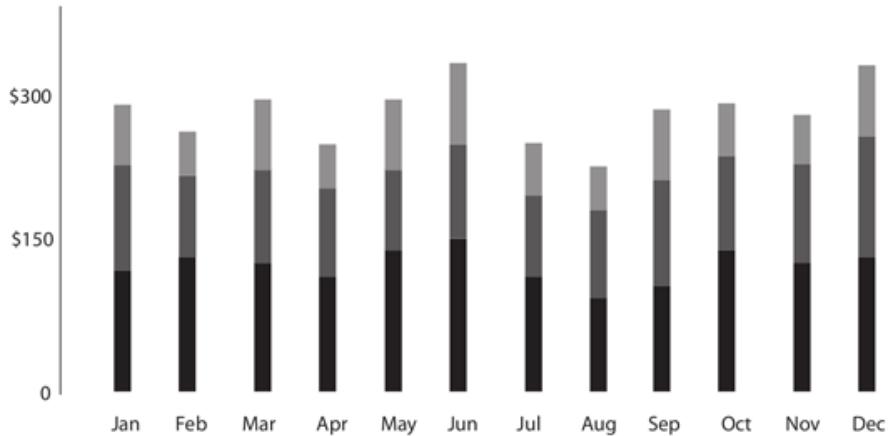
Sloupcový i čárový graf mají skupinovou variantu, kterou už se dají velmi pěkně vyprávět příběhy. Vezměme si například firmu se třemi pobočkami.



Obrázek 6: Skupinový sloupkový graf

Ted' 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. Když sloupce naskládáme nad sebe, vznikne takzvaný skládaný sloupkový graf, ve kterém už sice tak dobře nevidíme srovnání jednotlivých poboček mezi sebou, ale zase je jasnéjší, ve kterém měsíci nejvíce vydělává firma jako celek.

Oba grafy dívají 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.



Obrázek 7: Skládaný sloupcový graf

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 mapuobarvit, 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.

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 běžné. 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 nesejdete 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 zkousit 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 nenesete 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.

Datová žurnalistika v Česku

Zatímco předchozí kapitoly vycházely především z anglického originálu a prostředí, ta následující vám zprostředuje zkušenosti z datové žurnalistiky v České republice, jak ji vidí členové datového týmu [IHNED.cz](#).

Data v novinách: popis jednoho zápasu

O datové žurnalistice jsem poprvé veřejně mluvil na vsetínském Barcampu v lednu 2011. Byl jsem tehdy nadšený [webem mapy heren](#), který jednoduše a jasně ukazoval, jak zoufale jsou výsledky „regulace“ hazardu: devět z deseti hráčských doupat v centru Brna bylo v těsném sousedství školy nebo nemocnice. „Proč to neděláš?“ zeptal se mě po přednášce novinář Petr Kočí. Věděl jsem, že je to nesmysl. Česká média se horko těžko snaží přežít; o experiment, jakým byla datová žurnalistika, nemohla mít vážný zájem. Ještě komičtější byla představa, že bych se o to měl pokusit já.

O rok později jsme s Petrem v redakci IHNED zakládali datové oddělení. Začínal jsem s mírně nadprůměrnou znalostí Google Docs a obrovským respektem k vizualizacím Guardianu a New York Times. Petr na tom nebyl o moc líp. Bál jsem se, že jediné, co po nás zůstane, bude – rovněž obrovské – trapno.

Když novináři programují

V redakci v prvních měsících málokdo tušil, proč tam jsme: grafy přeci dělají grafici... Pak se nám ale podařilo dotáhnout do konce projekt, který smysl naší práce kolegům vysvětlil líp než deset školení. Jenomže [mapa odtahů aut v Praze](#) vznikala na pomery zpravodajského deníku neúměrně dlouho.

Už jen dostat se k policejním statistikám byla vrcholná diplomacie. Samotná vizualizace a vývoj webové aplikace nám pak coby neprogramátorem zabraly několik dalších týdnů. K umístění dat na mapu bylo potřeba sehnat síť pražských ulic, vymyslet logiku i uživatelské rozhraní aplikace, vylámat si zuby při pokusech o vytvoření vlastního funkčního kódu a pak sehnat vývojáře, který to celé poslepuje. Na testování moc času nezbylo. A o nějakém redakčním vývoji jsme si tehdy, před rokem, mohli nechat jen zdát.

V prvním plánu mapa odtahů odvyprávěla čtenáři jeho vlastní příběh: podíval se na ulice, kde obvykle parkuje, aby zjistil, zda mu nehrozí odtahy v míře větší než malé, a zkontoval si nejčastější příčiny odtahů ve své čtvrti či ulici.

Pak měla mapa ještě druhý plán. Vzorce, které se na mapě objevily až po chvíli hledání, a na nich postavené novinářské hypotézy: Ukázalo se, že nejvíce se neodtahuje tam, kde auta nejvíce překážejí, ale v blízkosti odtahových parkovišť, kde je to pro odtahovku nejjednodušší.

Každý, kdo pracoval v novinách, zná tu beznaděj, když se snaží politikovi či mocnému úředníkovi dostat pod kůži, ale respondent pokaždé vyklouzne. Když má ale novinář v ruce data, je to pro takového úhoře daleko těžší – viz [rozhovor kolegů Měřičky a Pavce](#) s bývalým šéfem městské policie Ludvíkem Klemou.

Pasti a pastičky

Mapa odtahů nás naučila dvě věci: za prvé, je důležité mluvit s redakcí. Za druhé, opravdu potřebujeme programátora. Nástroje, které jsou „zdarma každému dostupné na webu“ a „naučí se je používat každý“, jak hlásají propagátoři datové a občanské žurnalistiky, jsou sice fajn, ale pokud je nasadíte v ostrém redakčním provozu, narazíte brzy na různá omezení: proč Google Fusion Tables hlásí po nahrání dvou obsáhlějších tabulek překročení datového limitu? Proč jsou tak pomalé, často padají a jakmile se je naučíte používat, uživatelské rozhraní se změní? Proč když si tabulku z nich stáhnete ve formátu CSV a zkusíte nahrát do Google Spreadsheet, zjistíte, že si ani dva programy od jednoho výrobce mezi sebou nerozumí, protože každý používá jiné kódování češtiny? A proč programátoři dodnes nedokázali vyřešit formátování data a času nebo desetinné čárky tak, aby to fungovalo na celém světě tak dobře jako ve Spojených státech?

Pasti, které na datové žurnalisty číhají, nejsou jen technické, ale také byrokratické. Data produkovaná (nejen) českou státní správou jsou plná nepřesností a chyb, které strojové zpracování často znemožňují: Proč úředníci píší v jedné databázi nejdřív jméno, pak příjmení, a podruhé obráceně? Proč registr vozidel nesmyslně uvádí, že v celé zemi jezdí jen tři Škody Favorit? Proč úředníci místo jedničky píšou do identifikačního čísla firmy

malé l? (Odpověď: Protože se to tak naučili, když se lejstra vyplňovala ještě na psacím stroji.)

Jako zrušení cenzury

Novinář, který si tím vším projde, se nejspíš pokorně vrátí k tomu, že počítáč je opravdu jen lepší psací stroj. Nebo překoná přirozený odpor, který (jak si všímájí Rada a Žák v knize Študáci a kantoři) panuje mezi hmotou matematickou a hmotou lingvistickou již od prvohor, a rozhodne se sám programovat. Jen tak si může s daty dělat přesně to, co chce a potřebuje. A když se mu to daří, je to pocit osvobojující asi jako zrušení cenzury. *Programujte, nebo budete naprogramováni* – to je název knihy Douglase Rushkoffa, který se nám vybavil, když se nám podařilo spustit náš první funkční a pro noviny užitečný skript v programovacím jazyce R. Byl to scraper, který stahoval do přehledné tabulky výsledky loňských krajských voleb.

Jenomže dva neprogramátoři, kteří se učí programovat za pochodu v redakci zpravodajského webu a novin, které musí vyjít každý den, připomínají trochu zaměstnance stánku s rychlým občerstvením, kteří se jednoho dne rozhodnou, že budou místo párků v rohlíku servírovat hovězí po burgundsku. V receptu se píše, že se má dusit šest hodin, aby se rozsýpalo. Jenomže hrozí, že zákazníci mezitím odejdou a šéf je ze stánku vyhodí, když je maso ještě tvrdé.

Happy end?

Někdy v této fázi, v březnu 2013, nabrala naše cesta k jádru datové žurnalistiky nový, radostnější směr. Nejprve k nám přibyl třetí do party, mladý a nadšený datový novinář Honza Cibulka. A pak začalo v redakci pod vedením Martina Malého samostatné oddělení redakčních vývojářů. Kód, který jsme sami smolili několik týdnů, dokáže ostřílený programátor napsat za odpoledne, a na nás je, abychom skutečně hledali v datech téma a vymýšleli, jak je nejlépe odvyprávět. Místo technických problémů se můžeme víc soustředit na to žurnalistické: dolujeme z úřadů citlivá data (často přes zákon 106 o svobodném přístupu k informacím), víc o tématech mluvíme s redakcí, víc se snažíme v aplikaci „mluvit“ se čtenářem.

V naší práci přibyl jeden krok – kritické zkoumání už hotových aplikací a jejich ladění. (Tady se sluší poděkovat našemu největšímu kritikovi, Josefу Šlerkovi – kdykoliv máme pocit, že na aplikaci už není žádná práce, objeví se podvědomý kritik, „vnitřní Šlerka“, a vysvětlí nám, že to máme celé předělat.)

Doplňili jsme také svou definici datové žurnalistiky. Často jsme pomocí dat pouze „ilustrovali“ články jiných redaktorů. Obvykle potřebovali zaplnit chybějící místo mapou, vložit do článku tabulku nebo vytvořit graf. Tohle je práce, kterou se musí naučit sami redaktoři. „Opravdová“ datová žurnalistika by měla sloužit spíš investigaci než ilustraci; spíš objevování nových skutečností a provázaností, než jen popisování toho, co se „tak nějak vídá“.



Obrázek 8: Část datového týmu IHNED přebírá Cenu Googlu za inovativní profesionální online žurnalistiku za rok 2013 (autor: Richard Rameš)

Datová žurnalistika jako žánr časem zmizí. Stejně jako neexistuje „magnetofonová žurnalistika“ nebo „telefonní žurnalistika“, rozpustí se časem i datová žurnalistika v redakčním provozu. Stane se novinařinou, ke které není potřeba přidávat žádné přívlastky.

—Jan Boček a Petr Kočí

Přístup k informacím z pohledu uživatele

Následující text nemá ambice být univerzálním návodem, jak žádat státní správu o informace. Soudy, které v textu naleznete, lze přičítat jen autorově názoru a zkušenostem, které nemusejí být úplně přenosné. Nutno dodat, že pokud narazíte na osobu, která bude dostatečně autoritativně tvrdit, že v tomto textu je něco špatně, tak má pravděpodobně pravdu.

Infozákon a paní na ústředně

V novinářské praxi se jistě dostanete do situace, kdy budete od některé ze státních institucí (soukromých subjektů se to netýká, státních akciovek a s.r.o. ano) požadovat dokument či informaci, kterou nesdílejí sami a dobrovolně. Pokud si v tuhle chvíli vzpomeňnete na [zákon 106/1999 o svobodném přístupu k informacím](#), je to předčasné.

Než sáhnete po úředních postupech, ptejte se. Zní to dost banálně, ale má to svá úskalí. V první řadě se snažte vyhnout tiskovým mluvčím. Nejkvalitnější informace o datech, která se snažíte získat, vám poskytne člověk, který má dataset na starosti. Většina úřadů a institucí má na webu telefonní seznam. Začněte zde a v případě neúspěchu se vyptávejte na telefonní ústředně.

I když třeba nedostanete člověka, kterého potřebujete, jistě vám poradí, kdo má v podniku na starosti právě vaše téma. A pak zkuste znova ústřednu a tak dále. Mimochedem, telefonní operátoři a operátorky na vrátnicích institucí bývají hotový poklad. Většinou jsou na svém místě dost dlouho a mají přehled o chodu své instituce. Kromě znalostí telefonů často ví, čím se kdo zabývá a kdy bude k zastižení.

Pravdou je, že se instituce v Česku profesionalizují. V praxi to znamená, že jakmile utrousíte, že jste od novin, odkážou vás na tiskového mluvčího. I tak je dobré mít v předstihu osondováno, kdo má ve správě váš dataset a co obnáší jeho vydání. Můžete pak mluvčímu poskytnout přesné pokyny. Jinak máte slušnou šanci, že se do toho někdo z tiskového oddělení „kreativně“ vloží a celý proces poskytnutí dat zkomplikuje. Zdá se vám to absurdní? Tak například v pražském dopravním podniku o vydání datasetů novinářům rozhoduje oddělení marketingu a PR.

V novinářské praxi je takřka kritické získat data dohodou. Pokud dojde na přetahování prostřednictvím paragrafů (což je například zmíněný zákon 106), ocitnete se v zajetí lhůt a odvolacích procedur. Instituce má na vyřízení vašeho požadavku 15 dní (které si ale může dle libovůle až o 10 dní prodloužit).

Odvolání proti neposkytnutí informací se může vlekt měsíc a ani potom nemusíte svá data získat. Správní žaloba se pak počítá na měsíce až jednotky let.

Zajímejte se

Samozřejmě se běžně stává, že instituce odmítá data poskytnout. Často je za tím neopodstatněná obava z jejich použití, případně nějaký majetnický komplex konkrétního zaměstnance státní správy („my jsme ten dataset vytvořili, tak je náš, nikomu ho nedáme“). Rovněž se dotyčný pracovník může bránit přílivu další práce: „Váš dotaz je položen tak, že vyžaduje předzpracování či vytřídění dat.“ V takovém případě se můžete dohodnout na méně náročném řešení.

Když se instituce vydání dat brání, pomáhá vysvětlovat a zase vysvětlovat. Prokážte dobrou vůli; pokud je to možné, se správcem dat se sejděte, zkuste ho přesvědčit, že vás projekt za to stojí. Takoví lidé bývají perfektně seznámení s danou problematikou, často si užitečně rozšíříte obzory. Někdy ale sebevětší snaha nepomůže.

Když to jinak nejde

Pokud se instituce „šprajcne“ a vámi požadovaná data prohlásí za tajná, citlivá, interní, neveřejná (doplňte dle libosti), nezbyde než oficiální postup. Žádost podle zákona

106/1999 o svobodném přístupu k informacím musí obsahovat vaše identifikační a kontaktní údaje (stačí jméno, bydliště a datum narození), identifikaci subjektu, který o informace žádáte, a deklaraci, že jde o žádost podle zmíněného zákona 106/1999.

Rovněž nezapomeňte co nejjasněji uvést, o co vlastně žádáte. Nezapomeňte, že nemáte nárok na citlivá osobní data, věci kryté obchodním tajemstvím, informace o budoucích rozhodnutích a na data, které instituce nemá k dispozici a musela by je vytvořit (je toho ještě víc, podívejte se na výše uvedený zákon).

Pokud se nechcete formalitami zabývat, svěřte se do rukou **generátoru žádostí**, který vám žádost vygeneruje, odešle a pohlídá termíny. Instituce na vaši žádost musí odpovědět, pokud tak neučiní, můžete si stěžovat.

V ideálním případě vaši žádost vyřídí kladně a data dostanete. A když ne...

Bitva paragrafů

V případě, že žádosti není vyhověno, můžete se odvolat (ve lhůtě 15 dnů od obdržení rozhodnutí) či si stěžovat (do 30 dnů). Mimochodem, ono „nevyhovění“ má řadu podob: instituce například neodpoví, vaši žádost s nějakým odůvodněním zamítne (i jen částečně), nebo si naúčtuje nesmyslnou částku za vydání dat.

Konkrétní situace a jejich řešení najeznete na fóru webu [otevřete.cz](#). V diskusi zde občas odpovídá i Oldřich Kužílek, autor zákona 106. Než se ale zeptáte, zkuste nejdřív hledat. Rovněž zde najeznete [příručku](#) a novinky z oblasti přístupu k informacím, judikatura se v této oblasti stále vyvíjí. V zásadě ale platí, že pokud instituce vaši žádost zamítne, odvoláváte se. V ostatních případech (neodpoví, nastaví nesmyslnou cenu, popírá vlastnictví dat) si stěžujete.

Stížnost i odvolání musí obsahovat identifikaci rozhodnutí, které napadáte; stačí jednací číslo. V případě nečinnosti instituce odkazujete na předchozí žádost. Hezký diagram toho, jak žádost a další kroky probíhají, [najeznete na webu otevřete.cz](#).

U stížnosti napadáte procesní správnost rozhodnutí (dodržení termínů, formální nedostatky odpovědi atd.) či přiměřenost poplatků. Odvolání napadá konkrétní důvody zamítnutí žádosti; zde je vhodné zapárat v judikatuře k podobným případům a jimi argumentovat. A ano, tušíte správně, i dřívější rozsudky jsou k dispozici na webu [otevřete.cz](#).

O stížnosti či odvolání rozhoduje nadřízený orgán subjektu, od kterého data žádáte (například u pražské městské části je to magistrát). Ten může dřívější rozhodnutí zrušit a nebo potvrdit. V případě zrušení se žádost vrací k subjektu, od kterého žádáte a rozhoduje se o ní znova. Pravdou je, že instituce může tutéž žádost i napodruhé zamítnout a může tak učinit po každém dalším odvolání (které sice vyhrajete, ale data stále nemáte). A pozor, pokud subjekt nadřízený orgán nemá, rozhoduje o vašem odvolání sám.

Po vyčerpání opravných prostředků (odvolání či stížnost) se pak můžete obrátit na soud a podat správní žalobu. Zde už vás musí zastupovat advokát a na rozhodnutí můžete čekat několik měsíců nebo i let. Tomáš Němeček z Lidových novin takto například vymohl

výsledky státních maturit. Pravdou nicméně je, že shánět data k aktuálním tématům skrze soudní spory není moc reálné.

—Jan Cibulka

Doporučené nástroje a zdroje

Nástroje

- [Open Refine](#), open source aplikace pro čištění a strukturování dat.
- [Google Fusion Tables](#), dynamické tabulky umožňující kombinaci a sdílení dat z různých zdrojů a aplikací. Aplikace zvládá práci s daty do 100 MB a základní grafy a mapové vizualizace, dá se zvěnčí dotazovat pomocí API.
- [Tableau Public](#), software pro snadné vytváření interaktivních vizualizací. V základní verzi, tj. pro data do 100 tisíc řádků, je k dispozici zdarma; součástí webu jsou názorná instruktážní videa.
- [Datawrapper](#), open source nástroj pro tvorbu jednoduchých grafů.

Zdroje

- [Data Journalism Handbook](#), zhruba dvousetstránková původní verze příručky datové žurnalistiky, na které je tato publikace postavena. Webová verze je k dispozici zdarma.
- [DataDrivenJournalism.net](#), web mapující trendy v datové žurnalistice. Jeho součástí je komunitní mailinglist sdružující odborníky a nadšence z celého světa.
- [Záznam workshopu datové žurnalistiky](#) vedeného Karlem Minaříkem a Josefem Šlerkou.
- [School of Data](#), projekt OKFN zprostředkovávající návody pro základní i pokročilou práci s daty.
- [Datablog.cz](#), blog zaměřený na datovou žurnalistiku, otevřená data a kreativní práci s informacemi.
- [Otevřená data: příležitost pro Českou republiku](#), publikace OSF o otevřených datech a možnostech jejich využití.
- [Jak žádat o informace na základě zákona 106](#).

Tiráž

Základem tohoto textu byla publikace [Data Journalism Handbook](#), kterou přeložil, zkrátil, opět rozšířil a upravil [Tomáš Znamenáček](#) díky příspěvku od [Open Society Fund](#).

Text je zveřejněn pod licencí [Creative Commons Attribution+ShareAlike](#), což stručně řečeno znamená, že jej můžete libovolně šířit a dál na něm stavět, pokud uvedete odkaz na zdroj a výsledky své práce zveřejníte pod podobnou licencí. Zdrojový text publikace je na [GitHubu](#).

PDF verze je vysázena písmem [Skolar](#) Davida Březiny.