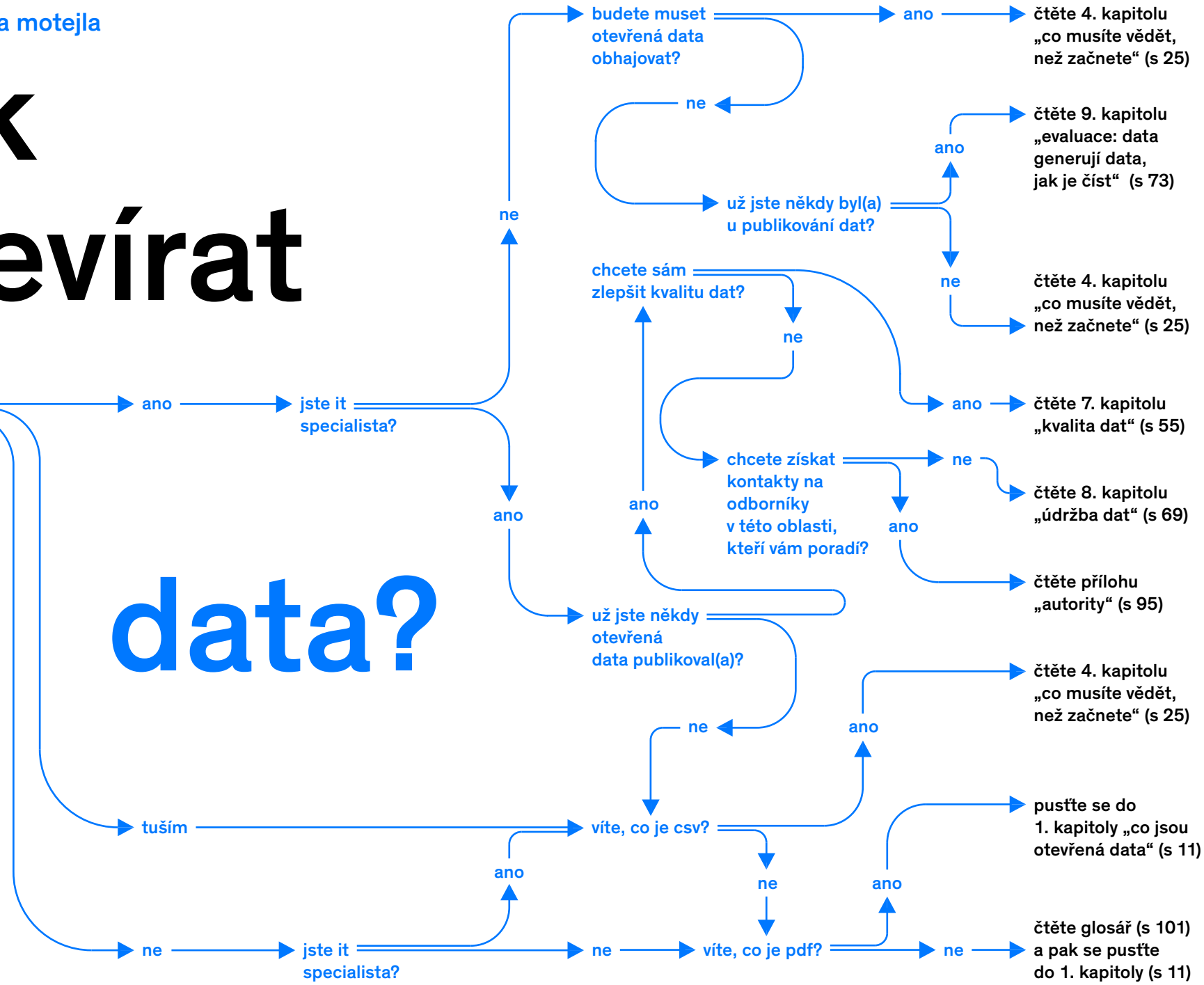


jak otevírat

víte,
co jsou
otevřená
data?

data?



čtete 4. kapitulu „co musíte vědět, než začnete“ (s 25)

čtete 9. kapitulu „evaluace: data generují data, jak je číst“ (s 73)

čtete 4. kapitulu „co musíte vědět, než začnete“ (s 25)

čtete 7. kapitulu „kvalita dat“ (s 55)

čtete 8. kapitulu „údržba dat“ (s 69)

čtete přílohu „authority“ (s 95)

čtete 4. kapitulu „co musíte vědět, než začnete“ (s 25)

pusťte se do 1. kapitoly „co jsou otevřená data“ (s 11)

čtete glosář (s 101) a pak se pusťte do 1. kapitoly (s 11)

obsah

●	předmluva	7
●	jak pracovat s touto příručkou	9
1.	co jsou otevřená data?	11
2.	proč otevírat data?	15
3.	otevřená data v česku: od vlády po města	21
4.	co musíte vědět, než začnete	25
●	mějte stále na paměti	26
●	naplánujte otevírání dat	27
●	nejčastější otázky a námitky (faq & fgo)	29
●	používejte vhodné formáty dat	33
●	přidejte data o datech: metadata	35
●	data publikujte surová a co nejpodrobnější	36
●	publikujte pod otevřenou licenci	36
5.	jaká data a kde je vzít?	39
●	zjistěte požadavky ze strany uživatelů	40
●	získejte data	41
6.	publikace dat	45
●	manuální a automatizované zveřejňování	46

Jak otevírat data?
Jakub Mráček a kolektiv

Vydal Fond Otakara Motejla v roce 2014

FOND OTAKARA MOTEJLA

Fond Otakara Motejla přispívá k přeměně veřejné správy na efektivní a transparentní službu občanům. Podporuje desítky nevládních organizací i malých lokálních občanských iniciativ a realizuje vlastní projekty, které zlepšují život v České republice.

Mezi jeho nejvýznamnější aktivity patří prosazování otevřených dat ve veřejné správě a vznik platformy online nástrojů pro aktivní občany NášStát.cz. Fond Otakara Motejla založila a spravuje Nadace Open Society Fund Praha a přispívají do něj soukromí dárci a firmy.

www.motejl.cz
www.otevrenadata.cz
www.nasstat.cz

Vznik manuálu podpořili účastníci Fóra INSPIRACE, které každoročně pořádá společnost ABRA Software na podporu podnikání. Děkujeme za podporu.



FÓRUM INSPIRACE

●	kam data fyzicky umístit?	47
●	způsoby zveřejňování	50
●	přidejte dokumentaci	51
●	zařad'te data do katalogu	52
7.	kvalita dat	55
●	pětihvězdičkové schéma	56
●	formáty dat v detailu	61
●	technická omezení na vlastním serveru	64
●	strojově čitelná a strukturovaná metadata	67
●	certifikát the open data institute	68
8.	údržba dat	69
●	co všechno údržba zahrnuje	70
●	náročnost a automatizace procesů	71
●	zapojte komunitu	71
9.	evaluace: data generují data, jak je číst?	73
●	co uživatelé doopravdy chtějí?	74
●	metrika: návštěvnost a počet stažení	74
●	co víte o svém uživateli?	75
●	co se s daty děje dál?	75
●	jak využití dat podporovat (a zpětně z toho profitovat)?	76
10.	kam dál: propojená data	77
●	kde a jak propojená data fungují?	78
●	další zdroje informací o propojených datech	79
A.	definice otevřených dat	81
B.	struktura dat	89
C.	autority	95
D.	výkladový slovník, glosář	101
E.	seznam zdrojů	109

předmluva

V roce 2009 svedlo několik britských autorit souboj s tamním ministerstvem dopravy. Šlo o jízdní řády, do té doby komerční data s drahou licencí. K soustředěnému tlaku na otevřenost úřadu se přidal i tehdejší premiér Gordon Brown. Ministerstvo brzy velkou část požadovaných dat zveřejnilo – zadarmo, aby je mohl využít kdokoliv s nápadem na inovaci. Vyvolalo to nejprve příval nových mobilních aplikací a vzápětí velký zájem o podobná data z jiných oblastí. Nakonec nový fenomén dostal i jméno: otevřená data.

Za minulých pět let přístup k datům státní správy zmasověl a demokratizoval se – otevřená data dávno nejsou záležitostí úzké skupiny počítačových nadšenců. Zasahují sféru politiky, byznysu, pracují s nimi vědci a studenti. Magazín [The Economist](#) přirovnal otevřená data k velkému třesku ve státní správě a odhadl roční přínos jen pro Evropu na 100 miliard eur. Otevřenost vládních databází vehementně podporují tři velké postavy mezinárodní politiky: americký prezident

Barack Obama, britský premiér David Cameron a místopředsedkyně Evropské komise pro digitální agendu Neelie Kroesová.

Donedávna byla otevřená data doménou vyspělých demokracií, jako jsou Velká Británie, Spojené státy a Švédsko. Rychle se k nim ale přidaly země, které v podpoře otevřenosti a rozvoje internetu spatřily šanci, jak se vymanit z temné minulosti: Estonsko, Keňa nebo Burkina Faso.

Teď se zájem o otevřená data probouzí i v Česku. Od roku 2011 otevírání dat prosazuje Fond Otakara Motejla v rámci svého úsilí o efektivní a transparentní veřejnou správu. Díky Fondu se pojem „otevřená data“ dostal do vládních dokumentů a postupně začaly data uvolňovat první instituce (Český statistický úřad nebo Česká obchodní inspekce) i města (například Děčín). Loni jsme vytvořili prostor pro konzultace: ve spolupráci s Vysokou školou ekonomickou a Matematicko-fyzikální fakultou Univerzity Karlovy vzniklo Fórum pro otevřená data. Na konci roku už ale byla poptávka tak vysoká, že nebylo v našich kapacitách ji uspokojit. Vydali jsme proto tuto příručku.

Věříme, že s naší příručkou bude zavádění otevřených dat zhruba stejnou výzvou jako smontování skřínky z IKEA: jednotlivé díly jsou k dispozici v podobě dostupných standardů; každá kapitola představuje jedno okénko návodu k montáži; a jak má skříňka vypadat, vám ukážeme pomocí úspěšných příkladů z tuzemských i zahraničních institucí, které už otevřená data zavedly. Těšíme se, že vedle několika průkopníků se do instalace otevřenosti pustí řada dalších úřadů.

Jakub Mráček, červen 2014

P. S. Podobně jako švédští nábytkáři vám ochotně odpovíme na každou tematickou otázku – jen místo ikony panáčka s telefonem hledejte kontakty na Fórum pro otevřená data v závěru knihy. Neodmyslitelným doplňkem knihy je také web otevrenadata.cz. Upozorní vás na aktuality z otevírání dat státní správy a přidá pár příkladů, které v příručce nenajdete.

jak pracovat s touto příručkou

Příručka chce být knihou, která uspokojí každého – politika, úředníka i programátora, začátečníka i experta na otevřená data. Nemusíte ji proto číst celou. Začněte třeba jednoduchým raz-dva-tři průvodcem, jak na otevřená data – najdete ho jen o několik odstavců dál. Ten vás pak přesměruje na kapitoly, kde se zvolené tematické věnujeme podrobněji.

Pro snazší orientaci navíc najdete na obálce příručky rozhodovací strom, který by vás měl navést na místo, kde se dozvíte přesně to, co právě potřebujete, na základě vašich současných znalostí.

1.

**co jsou
otevřená
data**

Než se pustíme do konkrétních postupů, jak data otevírat, vymežeme, co vlastně otevřená data jsou. Nejčastěji dostáváme tři otázky:

1 Z jaké oblasti data pocházejí?

Z jakékoli. Pojem otevřená data označuje formu, ne obsah. Může jít o data, která „produkuje“ každý z nás – týdenní záznamy tělesné hmotnosti, krevního tlaku nebo počtu uběhnutých kilometrů publikovaného aplikací RunKeeper [01]. Může jít o data, která získá univerzita ze svého výzkumu. Nejčastěji ale budeme v této knize mluvit o datech veřejné správy, tedy informacích, které sbírá nebo vytváří samospráva a státní správa.

2 O jaký formát dat se jedná?

Otevřená data jsou v nejširší možné definici data dostupná online. Může tedy jít o grafický formát (naskenovaný dokument, fotografie), geodata nebo textové dokumenty. Nejčastěji v této knize půjde o strukturovaná číselná data, tedy tabulky. Příkladem takových tabulek jsou třeba jízdní řády, seznam všech povolených předzahrádek ve městě, rozpočet města nebo statistika kriminality.

3 Jaké máme na otevřená data požadavky?

Pojem otevřená data je mladý a nespoutaný nějakou jednoznačnou definicí. Výčet požadavků se liší autor od autora – společně mají jen to, že otevřenost definují pomocí technických parametrů daného datasetu [02]. V příloze uvádíme šest konkrétních vymezení několika autorů, které doporučujeme pročíst.

⇒ Otevřená data nejčastěji chápeme jako soubory, které jsou dostupné online, jsou čitelná pro software a jsou co nejvíce strukturovaná. Přesné vymezení z několika pohledů najdete v příloze. ⇐

Pro podmínky Česka roku 2014 nicméně můžeme definovat otevřená data na základě dvou skupin požadavků:

- Česká státní správa a samospráva se musí řídit související legislativou, kterou tvoří zákony [106/1999 Sb.](#) a [123/1998 Sb.](#) a [Směrnice Evropského parlamentu a Rady č. 2003/98/ES](#) ze dne 17. listopadu 2003 o opakovaném použití informací veřejného sektoru.

- V Česku lze očekávat zakomponování vládních materiálů do prováděcích předpisů. Na mysl máme především metodiku publikace otevřených dat. Její doporučení vycházejí z trendů v zahraničí.

Sjednocením nároků z obou množin dostáváme standard otevřených dat v daném místě a čase. Tady je:

- Data jsou dostupná na internetu [03] a dohledatelná běžnými ICT nástroji a prostředky.
- Data jsou strojově čitelná, tedy ve formátu, který je jednoznačně čitelný nejen pro člověka, ale také pro počítačový program. Software potřebný ke zpracování takových datasetů by měl být volně (bezplatně) dostupný [04].
- Data jsou úplná, tedy zveřejněná v maximálním možném rozsahu.
- Data jsou dostupná s co nejmenšími překážkami (ať už technickými nebo legislativními) [05].

[1] V poslední době, zvláště díky masivnímu rozšíření smartphonů, je sběr dat o sobě samotných velkým trendem, který má vlastní jméno: [Quantified Self](#).

[2] V textu opakovaně užíváme termín dataset. Myslíme jím jakýkoli soubor obsahující data. Synonymem je třeba datová množina nebo soubor dat.

[3] Ve formulacích zákona 106/1999 Sb. jde o „způsob umožňující dálkový přístup“.

[4] Ve formulacích zákona 106/1999 Sb. „musí být zveřejněna i ve formátu, jehož specifikace je volně dostupná a použití uživatelem není omežováno“.

[5] Obvykle se žádají jasně definované podmínky užití (licence).

2.

**proč
otevírat
data?**

„Otevřená data jsou surovinou 21. století,“ píše bývalýředitel IT (Chief Information Officer) v Bílém domě **Vivek Kundra** [6]. Někdy se hovoří o „nové průmyslové revoluci“. Revoluce kolem otevřených dat se ovšem od té zmíněné v jednom liší: data nejsou surovinou, která by byla k dispozici pouze v omezeném množství a již by s každým použitím ubývalo.

Naopak. Otevřená data jsou dostupná všem bez rozdílu a každým použitím se jejich hodnota zvyšuje. Pouhým vystavením dat na internetu stát znásobuje jejich hodnotu. Nejzajímavější způsob využití vašich dat pravděpodobně vymyslí někdo jiný. Nelze dopředu odhadnout, jestli to bude malá, nebo velká firma, výzkumný tým, nebo státní instituce. S jistotou ale víme, že umístěním na web dostává soukromý i veřejný sektor příležitost data zhodnotit.

Čtyři důvody, proč publikovat data

1 Lepší služby

Jak data z Mapakriminality.cz zajímají policii

Michal Tošovský z centra Pro Police Otevřené společnosti se od roku 2003 snaží o otevřenější a s obyvateli lépe spolupracující strážce pořádku. Žádný z řady projektů, ať už šlo o informační portál, vydání gamebooku nebo e-learning pro samotné policisty, však nevedl k tomu, aby vyšší šarže projevíly o práci neziskovky vážný zájem.

V roce 2011 začal Michal pracovat na webovém projektu Mapakriminality.cz – sesbíral data o trestné činnosti z každého okresu a vytvořil jednotnou databázi se srozumitelnou mapovou aplikací. Projekt nejenže vyhrál prestižní soutěž datových vizualizací **On Think Tanks Data Visualisation Competition**, ale začal zajímat i policejní špičky: pročištěná data mohou sloužit jako nástroj řízení v celém sboru.

Najděte své předky

Ještě nedávno byla genealogie záležitostí několika málo lidí, kteří kromě schopnosti číst kurentem potřebovali také dostatek času, aby mohli objíždět matriky a archivy. Hledání rodinných kořenů je v současnosti velkým hitem, zcela jistě i díky ulehčení, které přinesly digitalizované

matriční knihy a služby jako Naše jména.cz a Kde jsme.cz. Ty stávají na datech ministerstva vnitra.

Vyberte pro své dítě nejlepší školu

Přáním každého, kdo respektuje svobodu jedince rozhodovat o vzdělání svých dětí, je, aby si mohl vybrat školu pro svého potomka. Výběr bez informací je losováním. V Nizozemsku proto na základě veřejně dostupných dat vytvořili **vyhledávač a srovnávač škol**.

2 Ekonomický potenciál

Hlídač změn ve firmách

V polovině května spouštěl Adam Kurzok svůj nový startup: hlídač změn v obchodních společnostech („Facebook pro firmy“), který umožní sledovat změny vlastnických struktur nebo akvizice. Projekt Daty.cz integruje data z mnoha veřejných rejstříků a nabízí je (od jistého počtu přístupů) jako placenou službu. Ekonomickou hodnotu dat lze těžko odhadnout, výsledky Daty.cz nám však mohou v budoucnu napovědět.

3 Transparentnost

Rozpočet každé obce na webu

Jak hospodaří i ta nejmenší vesnice Česka, ukazuje portál Rozpočet obce.cz. Data o hospodaření obcí, která v surové podobě zveřejňuje ministerstvo financí v **registru ÚFIS**, tu najdete v přehledné a porovnatelné podobě. Vedle nesporné hodnoty pro občany měst a obcí je tu ale také potenciál pro politologické a ekonomické analýzy, které v důsledku mohou pomoci efektivnější správě. Na Vysoké škole ekonomické už na základě dat z tohoto webu vznikly práce Vliv příjmové politiky na správu měst [7] nebo Výdaje na veřejnou správu měst [8].

[6]

Kundra, Vivek. Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect [online]. 2012 [cit. 2014-05-24]. Dostupné z: http://shorensteincenter.org/wp-content/uploads/2012/03/d70_kundra.pdf.

Srovnávače cen nad otevřenými daty

Jak postrčit lidi k lepšímu rozhodování třeba o tom, jaký fond penzijního připojištění nebo telefonního operátora si vybrat? Profesori Sunstein a Thaler [9], autoři proslulé knihy Nudge (Štouch), mají jasno – pomocí jejich přístupu RECAP: Record (zaznamenej), Evaluate (ohodnot) a Compare Alternative Prices (porovnej ostatní ceny). RECAP je díky otevřeným datům možný: už dnes existují srovnávače cen, které evidentně lepšímu rozhodování pomáhají.

Transparentní prodej veřejného majetku

Stát i samosprávy poměrně často prodávají majetek – třeba auta nebo nemovitosti. Vedle vlastních držav může jít třeba o exekuce. Účastníci veřejných dražeb takového majetku donedávna vyprávěli historiky připomínající divoký kapitalismus raných devadesátých let: o dražbách obvykle vědělo jen málo lidí, kteří ve spolupráci s úředníky kupovali pod cenou a pak se ziskem prodávali. Transparentnost do procesu vnesl až projekt Váš majetek.cz, který agreguje prodeje veřejného majetku. Díky němu se o něm dozví více zájemců a veřejná kasa má tak vyšší šance na příjmy, které odpovídají tržní hodnotě prodávané věci.

4 Efektivnější správa

Hodnota domovních adres v Dánsku

Noční můrou každého provozovatele e-shopu jsou neexistující adresy: zaslané zboží se vám vrátí zpět. Moderní elektronické obchody proto zavedly ověřování adres. K tomu je ovšem zapotřebí kompletního registru adres. V Česku to není problém. Adresní data byla k dispozici zdarma například v podobě **Územně identifikačního registru adres** (UIRADR). Dnes jsou ze zákona k dispozici přes **Registr územní identifikace a nemovitostí** (RÚIAN) ve strojově čitelné podobě. V Dánsku se ale za data o adresách až donedávna platilo. Díky tomu je možné porovnávat, jak se změnil počet uživatelů

před otevřením a po něm, případně odhadnout ekonomický přínos pro Dány [10] [11].

Období:	2004 až 2009	2010
Výnosy:	více než 60 mil. €	asi 14 mil. €
Náklady:	asi 2 mil. €	asi 0,2 mil. €
Návratnost investic:	22:1	70:1

[7] Zajiček, Petr. Vliv příjmové politiky města na jeho výdaje. (Diplomová práce) Praha, VŠE, 2013.

[8] Pířová, Tereza. Výdaje na veřejnou správu měst. (Diplomová práce) Praha, VŠE, 2013.

[9] Thaler, Richard H a Cass R Sunstein. Nudge (Štouch): Jak postrčit lidi k lepšímu rozhodování o zdraví, majetku a štěstí. Vyd. 1. Zlín: Kniha Zlín, 2010, 309 s. Tema (Kniha Zlín). ISBN 978-80-87162-66-8.

[10] Prosté pronásobení původní komerční ceny dat a rozdílu v počtu uživatelů samozřejmě nelze považovat za validní postup – snížení ceny (v tomto případě až na nulu) má podle zákonů trhu vliv na vzrůst poptávky. Ekonomická hodnota dat, která jsou omezeně dostupná a zcela otevřená, není jednoduše stejná. Výpočet přesto uvádíme, abychom zdůraznili, že otevřená data stimulují ekonomiku a mají obchodní potenciál.

[11] Shrnutí studie (v angličtině) najdete na bit.ly/VDAData.

3. I

**otevřená
data v česku:
od vlády
po města**

České republiky nejsou otevřená data veřejného sektoru úplně neznámá [12], zároveň ale nelze říct, že by státní správa vycházela zájemcům příliš vstříc. Většinu jejích dat zájemci stále dobývají buď složitým vytahováním informací z kódu stránky (tzv. scrapování) anebo formou žádosti na základě zákona o svobodném přístupu k informacím. Obě metody jsou velice náročné a finančně nákladné: úřady si za informaci leckdy účtují vysoké poplatky.

Z mnoha výstupů, které úřady publikují (například statistické ročenky), je nicméně často znát ochota data zpřístupňovat. Leckdy ovšem panuje jistá právní nejistota o možnosti například dalšího šíření dat a náročnosti transformace dat do využitelné podoby.

V září 2011 přistoupila Česká republika k Partnerství pro otevřené vládnutí a ve svém akčním plánu se zavázala zpřístupnit desítku datasetů veřejné správy. Do června 2012 však byly v plné míře zpřístupněny pouze **výsledky voleb** spravované Českým statistickým úřadem, a data-sety z **Monitoru státní pokladny** [13]. Přestože v únoru 2013 Fond Otakara Motejla společně s více než 50 firmami a přes 700 jednotlivců vyzval vládu k naplnění závazků, zbylé datasety stát neotevřel. V nejbližší době vládu navíc stejně jako zbytek Evropské unie čeká úprava legislativy. Do července 2015 je třeba do českých zákonů integrovat **evropskou směrnicí 2013/37/EU** o opakovaném použití informací veřejného sektoru („PSI“ [14]).

Některé úřady mezitím začaly zpřístupňovat data dobrovolně. Česká obchodní inspekce publikuje **otevřená data o provedených kontrolách**, jejich výsledcích a udělených sankcích; Český telekomunikační úřad zpracoval komplexní analýzu svých dat a **některá z nich už publikoval**. Otevírání dat ve spolupráci s Fondem přislíbily kraje i města, mezi prvními byly Děčín, Opava a Kuřim. Praktický potenciál otevřených dat pak ukázal první ročník soutěže **Společně otevíráme data**, který jsme uspořádali v roce 2013. Ze soutěže vzešla řada užitečných aplikací v oblasti veřejné dopravy, zdravotnictví nebo transparentnosti politiky. Tím však práce zdaleka nekončí. Aby otevřená data měla podobný ekonomický a společenský dopad, jako je tomu třeba ve Velké Británii, je třeba publikovat výrazně větší množství dat.

[12] Komplexní zprávu o otevřených datech v České republice poskytuje Dušan Chlapek, Jan Kučera, Martin Nečaský a Michal Kubáň. Open data and PSI in the Czech Republic. [online]. 2014 [cit. 2014-05-24]. Dostupné z: <http://www.epsiplatform.eu/content/open-data-and-psi-czech-republic>.

[13] Konkrétně jde o finanční statistiky (státní dluh, vládní finanční statistika) a účetní záznamy a finanční údaje z CSÚIS. Dostupné z: <http://www.epsiplatform.eu/content/open-data-and-psi-czech-republic>.

[14] Směrnicí z roku 2013 se mění směrnice 2003/98/ES o opakovaném použití informací veřejného sektoru, o níž hovoříme v příloze.

4.

**co musíte
vědět, než
začnete**

Než se pustíte do pátrání po datech ke zveřejnění, ještě se na chvíli zastavte. V následující kapitole vám přiblížíme, co vás při otevírání dat čeká a na co je dobré se připravit.

Jedním dechem dodáváme, že publikovat otevřená data není žádná věda. Upřímně, zvládnete to, pokud umíte editovat jakýkoli web. Zároveň ale není třeba, abyste prošlapávali stokrát prošlapané cesty; proto popisujeme české i zahraniční, dobré i špatné zkušenosti.

↳ Česká obchodní inspekce:

Tiskový mluvčí inspekce Martin Tajtl na konferenci ISSS v dubnu 2014 sdílel svou zkušenost z úspěšné publikace otevřených dat ČOI. Práce mu zabrala méně než tři pracovní dny. Podobně v Děčíně publikace prvních datasetů proběhla během dvou týdnů. ←

Mějte stále na paměti

Hned na začátku je důležité zmínit dvě základní pravidla, která by měla být při zpřístupňování dat dodržována [15]:

Začněte s málem

Pokud se chystáte zpřístupnit data, začněte s malými objemy, které bude možné jednoduše a rychle zveřejnit. Není nutné publikovat veškerá dostupná data okamžitě po spuštění celého projektu (je ale pravda, že čím více dat se vám zpočátku podaří zveřejnit, tím lépe).

Zveřejňujte data brzy a často

Poskytujte současným i potenciálním uživatelům co nejčerstvější data, výrazně to zvýší jejich relevanci.

Je také třeba si uvědomit, že většina zpřístupněných informací nebude použita koncovými uživateli, ale spíše zprostředkovateli dat. Ti je zpracují do srozumitelné podoby a následně poskytnou koncovým uživatelům.

Naplánujte otevírání dat

Otevírání dat v úřadu by v žádném případě neměla být hurá akce. Popisovaný návod se může oprávněně zdát jako zdoluhavý – je to způsobeno množstvím hráčů, kteří do zveřejňování dat promlouvají.

1 Zmapujte hráče

Každá instituce má vedle oficiální hierarchie a rozdělení kompetencí i své šedé eminence, poradce s neformálním vlivem nebo doyen, které všichni respektují. Předtím, než se záměrem otevírat data vyrukujete na nejvyšších úrovních, všechny tyto hráče si zmapujte. Kdo může mít dané téma v kompetenci? Čí spolupráci budete potřebovat? Kdo bude projektu bránit a proč?

2 Najděte průsečíky zájmů

S mapou hráčů budeme dále pracovat: promýšlejte všechny možné interakce a hledejte, kde existují (nebo mohou existovat) průsečíky mezi zájmy jednotlivých lidí. Typickými křížovatkami mohou být:

- V úřadu existuje komise pro transparentnost nebo komise pro spolupráci s komerční sférou.
- Tiskový odbor chystá zásadní rekonstrukci webových stránek.
- Některý odbor pořizuje nový software.

3 Získejte spojení

Snažte se v mapě hráčů najít místa, kde očekáváte největší podporu (ideálně win-win situace). Sejděte se s nimi, projekt jim detailně vysvětlete, proberte, čím můžete být užiteční. Časová investice do hledání spojenců se záhy vyplatí: podpoří vás v jednání s těmi, kdo rozhodují,

[15]

Pasáž je překladem části The Open Data Handbook – Open Knowledge. The Open Data Handbook [online]. 2011 [cit. 2014-05-22]. Dostupné z: <http://opendatahandbook.org>.

budou základem pracovní skupiny a poskytnou vám nejrychlejší a nejcitlivější zpětnou vazbu.

4 Oslovte decision makery

S několika (pokud možno vlivnými) spojenci v zádech je čas na audienci na nejvyšších místech. Politickou podporu považujeme za klíčovou, mimo jiné i proto, že brzy budete potřebovat spolupráci na úrovni velkého množství odborů, jejichž vedení nemusí k vašemu projektu chovat sympatie. Politická záštita pomůže: otevírání dat přestane být myslích kolegů projektem z jedné kanceláře, ale stane se šéfovým přáním.

V této fázi můžete narazit na odpor. V další podkapitole najdete nejčastější argumenty, proč to nejde, a odpovědi na ně.

5 Vytvořte pracovní skupinu

Co nejrychleji vytvořte pracovní skupinu angažující všechny relevantní hráče. Zvláště rychle zapojte ty, kteří mají k projektu nejvíce výhod.

Typicky budou členy takové pracovní skupiny:

- Tiskový mluvčí
- Vedoucí IT oddělení
- Tajemník/ředitel úřadu
- Vedoucí odborů, které budou data publikovat
- Právní odbor

Pracovní skupina musí odpovědět na několik klíčových otázek:

- Jaké datasety by stálo za to zveřejnit? (Pokud netušíte, kde je hledat, podívejte se do kapitoly Jaká data a kde je vzít?)
- Které datasety je vedení úřadu ochotno zveřejnit?

- Jaká data chtějí budoucí uživatelé? (Víc v podkapitole Požadavky ze strany uživatelů.)
- Jaké jsou konkrétní náklady a přínosy? (Podkapitola Analýza nákladů a přínosů.)

6 Sestavte harmonogram

Od vytvoření pracovní skupiny do publikace prvních datasetů by nemělo (ani na velkém úřadu) uplynout více než dva tři měsíce. Mějte tento milník stále před sebou a snažte se ho stihnout.

7 Spusťte to

Vše, co jste si naplánovali, teď čeká na realizaci.

8 Pochlubte se

Máte za sebou úspěšný projekt. Má-li se na něj někdy navázat, je třeba slíznout smetanu a pochlubit se. V další vlně zveřejňování bude všechno snazší.

9 Poučte se

Promýšlejte, co dělat příště lépe. Věnujeme tomu celou kapitolu Evaluace: Kterak data generují data a jak je číst.

Nejčastější otázky a námítky (FAQ & FGO)

Argumenty, proč to nejde, jsou napříč úřady velmi podobné. Je tedy pravděpodobné, že na ně narazíte i vy. Pro vaše pohodlí jsme na ně rovnou připravili i odpovědi.

Přestože se za hlavní překážku publikace otevřených dat nejčastěji považují finanční náklady, ve skutečnosti je to lidský faktor.

Konkrétní důvody jsou různé: neporozumění, v čem jsou otevřená data přínosná, nechuť k přílišné otevřenosti úřadu nebo apriorní odmítání internetových fenoménů.

Otevírání dat je konfliktní téma. Je proto třeba, aby mělo záštitu – z čím vyššího místa, tím lépe. Velká Británie se stala světovou jedničkou v otevřených datech díky šťastné souhře mnoha okolností, tou klíčovou ale byla **osobní opakovaná podpora premiéra Davida Camerona**.

Pokud jste ve vedení úřadu, který chcete otevřít, osobně na proces otevírání dohlížejte. Pokud jste řadovým zaměstnancem úřadu, hledejte pro projekt podporu na vyšších pozicích.

Ted' už seznam typických námitek

1 Vlastně to není moc zajímavé.

Záleží na úhlu pohledu. Přínosy, které vidíme v zahraničí (viz kapitola Proč otevírat data?), nám přijdou poměrně úctyhodné.

2 Je to příliš složité.

Není, zvládají to i malé obce a úřady. Jde jen o to, jaký rozsah a způsob publikace dat si vyberete.

3 Papír je papír...

V 21. století na takovou námitku těžko reagovat jinak než smíchem. Navrhujeme oslovit všechny zájemce o data dotyčného úřadu, zda by si přišli stoupnout do fronty před jeho dveře s požadavkem na vytištěná data.

4 Otevřená data jsou nová forma komunismu: datasety jsou surovinou, mají hodnotu, ať si za ně zájemci zaplatí, pokud je potřebují.

Za data jsme již jednou zaplatili, a to v podobě daní. NASA kdysi prodávala snímky z Hubbleova teleskopu, než soud jasně řekl, že NASA je placena z veřejných peněz, a tak i její výstupy mají povahu veřejného statku [16].

5 Pustit z ruky neinterpretovaná data znamená, že jich někdo zneužije: špatně je vyloží, způsobí tím škodu a ta bude vymáhána na nás.

Pokud k datům vytvoříte dokumentaci, v níž popíšete, co znamenají a co je korektní s nimi dělat, nemusíte se bát – ani žaloby za způsobené škody, ani ztráty reputace. Podle dosavadních zkušeností k takovým dezinterpretacím prakticky nedochází.

6 Zahrneme data do většího projektu, vytržená publikace dat je nesystémová jednotlivost.

S tím je třeba souhlasit (proč by každá obec měla například publikovat otevřená data o rozpočtech, když už je publikuje ministerstvo financí za všechny?). Vždy je ale třeba ostražitě hlídat, jak ten větší projekt vypadá. Setkali jsme se s nesourodými balíčky „otevírání radnic“, ve kterých se dodavatelská firma věnovala optimalizaci procesů, interakci občan-úředník, digitalizaci dokumentů a mimochodem i takzvaným otevřeným datům...

7 Naše data jsou v nevyhovující kvalitě.

Alespoň tedy dejte vědět, že je máte. Pokud uživatel neví, co hledá, nežádá to. V případě, že o data bude zájem, možná se vyplatí do zlepšení jejich kvality investovat (více v kapitole Kvalita dat).

[16] U některých institucí může přesto jít o seriózní problém, způsobený faktem, že si podle zákona nebo zřizovací listiny na sebe musí vydělávat. Příkladem je Český úřad zeměměřický a katastrální.

[17] Tajtl, Martin. Otevřená data o kontrolách ČOI. In: [online]. 2014 [cit. 2014-05-24]. Dostupné z: http://www.issc.cz/archiv/2014/download/prezentace/coi_tajtl.pdf. Přebíráme námitku z anonymního článku v magazínu Smart Cities. Zdrojový článek je: Informační dopravní systém versus otevřená data. Smart Cities. 2014, 1., č. 1, 28–33.

8 Zahltí nás to.

Záleží na tom, jak si nastavíte procesy a priority. Při správném plánování je zátěž nízká (rozhodně v poměru k potenciálním ziskům). Například mluvčí České obchodní inspekce [17] říká o zkušenosti se zveřejňováním toto:

- Spuštění včetně analýzy trvá 10–20 hodin.
- Provoz zabere asi 15 minut na aktualizaci, při čtvrtletní aktualizaci je to hodina za rok.
- Rozvoj a zdokonalování spolkne zhruba 8 hodin ročně.

9 Právníci chtějí vlastní licenci.

Jejich výklad je minoritní. Nejspíš je k tomu vede některá z výše uvedených obav. Licence Creative Commons 4.0, kterou zmiňujeme v podkapitole Publikujte bez právních omezení, pasuje na otevřená data výborně. Ve zmíněné kapitole najdete i odkaz na blog, kde právníci právníkům všechno vysvětlují.

10 Otevřená data jsou nástroj šikany [18].

Ano, žádají po úřadech práci, která k výkonu veřejné moci není přímo potřeba. Občané ale mají na informace ze zákona nárok; otevřená data navíc mohou v důsledku úředníkům práci ušetřit.

11 Některé datasey nelze publikovat: jsou to příliš velké nebo složité soubory, které není v technických možnostech úřadu smysluplně publikovat.

To může být skutečný problém – našťastí snadno řešitelný. Data prostě nebudete mít na svém serveru, ale v cloudovém úložišti (více v podkapitole Kam data fyzicky umístit).

12 Informace obsažené ve smlouvách, fakturách nebo interních dokumentech jsou z mnoha důvodů nezveřejnitelné – obsahují osobní údaje, obchodní tajemství nebo zneužitelné informace.

Samozřejmě – údaje, které jsou podle zákona neveřejné, nám ze seznamu datasetů vypadnou [19]. Anebo provedete jejich anonymizaci. Často se ovšem tenhle argument používá i pro datasey, kterých se netýká.

Anonymizací myslíme odstranění těch sloupců tabulky, které jednoznačně identifikují nějakou osobu nebo subjekt (IČO, rodné číslo, bankovní spojení, kombinace jména a adresy atd.). Ne vždy je to nutné; pro každý případ je třeba nastudovat příslušnou legislativu, zvláště **Zákon o ochraně osobních údajů 101/2000 Sb.** Jeho výkladem je pověřen Úřad pro ochranu osobních údajů Uoou.cz, se kterým je vhodné se o sporných případech poradit.

Používejte vhodné formáty dat

V úvodu jsme se zmínili, že definice otevřených dat úzce souvisí s jejich formátem. Formát totiž určuje, jakým způsobem lze s daty zacházet. Některé softwarové formáty, jako PDF nebo JPEG, jsou zcela nevhodné.

Proč?

Data musí být strojově čitelná

Naskenovaný dokument je dobře čitelný pro člověka, nikoli však pro stroj. Zatímco čtenáři je celkem jedno, jestli se dívá na znak A v textovém dokumentu, nebo obrázek písmene A, počítač si druhou variantu s písmenem nespojí. Strojovou čitelností rozumíme právě to: data mají podobu digitálních znaků, ne obrázku. To je problém formátů určených primárně pro záznam grafiky nebo fotek, jako je PDF.

Metod, jak převést data z „papíru“ do digitální podoby, je víc. Nejčastěji slouží k převodu obrázku na digitální text program typu OCR (Optical Character Recognition). Než se ale do digitalizace pustíte,

[18] Další informace: Mráček, Jakub. Kdo vydělává na monopolu na státní data?. Lupa.cz [online]. 2013 [cit. 2014-05-24]. Dostupné z: <http://www.lupa.cz/clanky/jakub-mracek-kdo-vydela-na-monopolu-na-statni-data>.

[19] Co je a není neveřejné (nebo dokonce tajné), dobře vysvětluje zákon 106/1999 Sb.

zvažte (ideálně formou analýzy nákladů a přínosů), zda se taková práce vyplatí. Začínat je lepší s tím, co už elektronicky máte.

↳ Příklad zveřejňování smluv:

Řada obcí zveřejňuje na webových stránkách úřadu smlouvy jako naskenovaná PDF. To je jistě chvályhodné, ke strojovému zpracování se ale tenhle formát nehodí. Není-li možné mít smlouvy strojově čitelné celé, pak je třeba do databáze vypsát alespoň identifikační údaje (číslo smlouvy, IČO smluvních subjektů, předmět, data a částky). Jak správně zveřejňovat smlouvy, ukážeme v příloze. ←

Preferujte otevřené formáty

I když jsou informace zveřejňovány elektronicky na internetu ve strojově čitelném formátu, mohou být k ničemu kvůli uzavřenosti formátu. O co jde: otevřeným formátem se myslí takový, ke kterému je k dispozici dokumentace zdarma. Tato dokumentace musí být volně použitelná, bez restrikcí vynucených zákony o duševním vlastnictví. Naopak „uzavřený“ je formát, který nemá zveřejněnou dokumentaci nebo kde je z důvodu ochrany duševního vlastnictví jeho použití omezeno. Zveřejnění informací v uzavřeném formátu může způsobit výrazné překážky pro jejich další využití, neboť nutí uživatele zakoupit si příslušný software.

Výhodou otevřených formátů je to, že umožňují vývojářům volně vytvářet k formátům různé programy a služby, což minimalizuje překážky v dalším použití informací obsažených v daných formátech.

Používání specifického formátu, pro který není k dispozici veřejně přístupná dokumentace, může způsobit závislost na softwaru třetí strany nebo vlastnicích licencí formátů. Pokud přesáhnou náklady na přístup k formátu únosnou mez, může v nejhorším případě dojít k tomu, že zveřejněné informace nebude nikdo používat.

Nejvhodnější formáty

V téhle kapitole se nebudeme pouštět do detailů, pouze vyjmenujeme doporučené formáty. Jejich podrobný popis, včetně zdůvodnění, proč jsou vhodné, najdete v kapitole Publikace dat. Ideální formáty jsou JSON, XML nebo RDF. Stačí ale i tabulky ve formátu CSV (je lepší než uzavřený XLS). Pro geodata doporučujeme GeoJSON.

Přidejte data o datech: metadata

Metadata nejsou nic jiného než popisky datasetu. Plní stejnou funkci jako štítky na lahvích ve spíži – bez nich sice skrze stěny vidíte, že se jedná o mouku, stěží už ale poznáte, jestli jde o polohrubou, či hrubou, a už vůbec nevíte, kdy jste si ji koupili. Malý štítek na sklenici to spraví – pomůže vám se ve spíži orientovat.

Nejjednodušší formou je na stránku, odkud je možné data stáhnout (může jít o záznam v katalogu otevřených dat), umístit tabulku s takovými popisky, jako to dělá [Český telekomunikační úřad](#). Struktura a obsah metadat jsou samostatným tématem, u vybraných datasetů se je snažíme řešit v příloze.

Obecně jsou požadavky stejné jako na samotné datasety:

- Standardizovaná a strukturovaná metadata (ideálně s využitím slovníku [DCAT](#))
- Strojově čitelná metadata
- Stažitelná metadata

Obsahově by u každého datasetu měl být uveden alespoň smysluplný název, stručný popis, datum publikace a kontakt na správce dat (zárodky takových metadat můžeme vidět třeba [na webu města Děčín](#)).

Geografická data mohou být popsána metadaty podle některé z norem, nejlépe ISO 19115. Metadatový záznam lze vytvořit ručně z formuláře [na stránkách národního geoportálu INSPIRE](#) nebo v některém jiném online generátoru. Vzápětí jej můžete zaregistrovat v některém z metadatových katalogů, ať už na národním geoportálu INSPIRE nebo v katalogu domovské organizace. Katalogové služby spolu navzájem „mluví“ díky standardu OGC Catalogue Service for Web (CSW), a tak lze vyhledávat metadatové záznamy přes několik serverů. Na závěr můžete díky metadatům prolinkovat datasety podle příslušnosti a hierarchie.

Data publikujte surová a co nejpodrobnější

Meteorologické stanice měří teplotu vzduchu v pravidelných intervalech a vyprodukují tak velký objem dat. Ta slouží k upřesňování matematických modelů, kterými se vědci snaží vývoj počasí předvídat. V televizním zpravodajství se však vždy objeví jen denní a noční průměr, tedy pouze dvě souhrnné hodnoty.

Zatímco z původních naměřených dat je možné průměr kdykoli spočítat, obráceně to nejde. Vyhodnocením, shrnutím nebo interpretací dat se ztrácí informace. Pro opětovné využití uživatelé uvítají, pokud poskytnete data v původní podobě, maximálně detailně [20] (např. hodinové koncentrace škodlivin v ovzduší namísto denního průměru).

Výhodou pro poskytovatele dat je nakonec i to, že žádaný přístup znamená méně práce – dataset prostě jen umístíte na web. Ideální samozřejmě je, pokud k němu připojíte i dokumentaci.

Argumentem proti takovému postupu často bývá obava z dezinterpretace. Pokud je to i váš problém, přeskočte na oddíl Nejčastější otázky a námítky.

Anonymizace dat

Skutečným problémem je ochrana osobních údajů. Zveřejňujete-li například zápisy z jednání zastupitelstva, výši platů nebo smlouvy, pravděpodobně budete muset řešit tzv. anonymizaci. U dokumentů je třeba osobní údaje vymazat, v případě databázi je třeba odstranit sloupec s jednoznačnými identifikátory [21].

Publikujte pod otevřenou licenci

Zveřejněná data jsou k ničemu, pokud nedáte zájemci svolení k jejich použití. Abyste zamezili nedorozuměním a nejistotám, je dobré podmínky užití (anebo licenci, chcete-li) explicitně zveřejnit (ideálně jako metadata).

Licenci pro otevřená data existuje celá řada, odborníci se ale poslední dobou shodují především na Creative Commons 4.0. Přestože tato verze ještě není přeložena do češtiny, lze ji v Česku používat. O překlad se stará nevládní organizace Iuridicum Remedium (iure.org) ve spolupráci s [Ústavem práva a technologií Právnické fakulty Masarykovy univerzity](#) (otevřené licenci se věnují i [na svém blogu](#)). Dokončení překladu očekáváme v průběhu roku 2014.

Srovnání Creative Commons s ostatními dostupnými licencemi najdete v brožuře [Veřejné licence v České republice](#).

↳ Jak použít Creative Commons 4.0:

- Publikujte data (více v oddílu Publikace).
- Vyplňte „bibliografické údaje“ na stránce [Creative Commons](#).
- Nezakazujte možnost upravovat váš dataset.
- Dovolte ho použít pro komerční účely.

Do metadat na webové stránce, odkud bude možné data stáhnout, umístěte též položku Licence a vložte HTML kód, který vám výše uvedený formulář vygeneroval (obsahuje sémantické informace v podobě RDFa). <←

[20] Odborný termín pro takovou míru detailu je „granularita“. Český ekvivalent „zrnitost“ se neuzívá.

[21] Existují ale i jiné situace, kdy primární data publikovat nelze. Například Český statistický úřad má nařízenou ochranu subjektů statistických šetření.

5.

**jaká data
a kde je vzít?**

Po rozhodnutí publikovat data obvykle přicházejí velké rozpaky: která data tedy zveřejníme? Odpověď by měla být průnikem dat, které si jejich budoucí uživatelé přejí (nechcete přece publikovat data, která se budou bez povšimnutí válet na disku) a jejichž zveřejnění nebude příliš drahé.

↳ Dobrý příklad Českého telekomunikačního úřadu: ČTÚ ohlásil záměr otevřít svá data v listopadu 2013. Výrazný vliv na rozhodnutí měl předseda rady úřadu. Zadal nejprve provedení analýzy, které datasety by úřad mohl otevřít. Z ní vzešla asi padesátka datasetů. Deset z nich bylo označeno jako prioritní a zveřejněny byly v březnu 2014. Na dalších 18 měsíců jsou naplánovány další dvě vlny publikování. Vybrané datasety odpovídaly uvedeným předpokladům – předseda rady u nich očekával praktické využití a současný stav dat nevyžadoval zásadní úpravy. ←

Zjistěte požadavky ze strany uživatelů

Data nepublikujete proto, abyste si mohli odškrtnout, že jste i letos udělali něco pro transparentnost, ale prostě proto, aby je někdo využil – jediné tak získají přidanou hodnotu.

Nejlepším začátkem je prostě se uživatelů zeptat. Pokud na to nemáte čas nebo nemáte vyhraněnou komunitu potenciálních uživatelů, vybírejte podle následujících doporučení:

- Zjistěte, jaká data jsou často žádána podle zákona 106/1999.
- Zjistěte, co publikují instituce podobné té vaší.
- Vysokou míru využitelnosti mají ekonomická data nebo geodata.

↳ Jak zjistit, co si lidé přejí?

● Vytvořte formulář a propagujte ho na sociálních sítích: příkladem je „[seznam datových přání](#)“ Fóra pro otevřená data.

● Ptejte se vývojářské komunity: existuje řada vývojářských fór, specializovaných webů nebo mailových konferencí. Ptejte se elektronicky nebo je pozvěte na pracovní snídani.

● Využijte zkušeností ostatních: nejjednodušší cestou je porozhlédnout se po internetu. Pokud v podobné instituci zabodovala nějaká data, můžete čekat podobnou poptávku i ve svém případě. ←

Vytvořte analýzu nákladů a přínosů

Analýza nákladů a přínosů (anglicky Cost-Benefit Analysis, CBA) je technika pro zhodnocení efektivity projektu ještě před tím, než se pustíte do jeho realizace. Rozšiřuje čistě ekonomickou analýzu o společenské dopady projektu (externality). Pokud by kompletace jednoho datasetu znamenala měsíční práci jednoho úředníka, pravděpodobně se nevyplatí data zveřejňovat. Pokud ovšem víte, že o dataset stojí místní podnikatel, který nad ním postaví službu lákající občany na váš web, stojí taková investice za úvahu.

O metodice analýzy nákladů a přínosů existuje celá řada zdrojů, doporučujeme některé z nich prostudovat [22].

Získejte data

Z průzkumu potřeb budoucích uživatelů pravděpodobně vyplyne zájem o data, která nejsou pro otvírání zcela připravena – bude třeba je teprve získat a převést do podoby alespoň „tříhvězdičkových“ dat (viz kapitola Kvalita dat). Jak a kde data vzít?

↳ Data je téměř vždy lepší zveřejnit hned a v horší kvalitě než nechat uživatele dlouho čekat. Zda má smysl investovat do vyšší kvality dat, ukáže poptávka [23]. ←

[22] Analýze veřejně prospěšných investičních projektů se věnuje disertační práce Patrika Siebera – Sieber. Stanovení hodnoty veřejně prospěšných projektů – Cost-Benefit Analysis. Praha, 2005. Dostupné z: <https://webhosting.vse.cz/ekisl/prace/Sieber.pdf>. Autoreferát k doktorské disertační práci. Vysoká škola ekonomická v Praze.

[23] Tradiční námitkou je, že nízká kvalita dat potenciální uživatele odradí. Deficit v kvalitě tedy dorovnávejte otevřeností komunikace: datasety, o jejichž nízké kvalitě víte, označte jako nekalitní a žádejte zpětnou vazbu, zda má cenu na kvalitě pracovat.

Používáte DMS nebo CMS? Spoustu dat už máte

Instituce využívající systém na správu oběhu dokumentů (DMS) mají jistou výhodu – rychle v nich zjistíte, jakými daty úřad disponuje. Některé systémy jsou navíc propojeny se správou obsahu webu a dokážou dokumenty rovnou na webu publikovat. Správce systému vám prozradí, kolik toho za vás může vykonat software.

Typickou situací je, že data shromažďujete v nějakém databázovém softwaru (může jít třeba o účetní systémy, různé evidence nebo nástroje spisové služby). Obvykle takové programy nabízejí exporty do různých formátů. Které zvolit, zjistíte v kapitole Publikace dat.

Digitalizace dat

Hodně (jak historických, tak aktuálních) atraktivních dat mají úřady pouze v papírové podobě. Dokumenty je možné digitalizovat automatizovaným skenováním. Pokud se do toho pustíte, mějte na paměti, že:

- Vstupní náklady na skenování jsou drahé. Nejde jen o multifunkční zařízení nebo v lepším případě automatický skener, který možná už máte, ale především o lidské kapacity – někdo musí naskenované soubory třídit, katalogizovat apod.

- Jednorázové skenování starších dokumentů je odůvodnitelné.

- Namísto skenování nově vznikajících dokumentů se raději pokuste nastavit procesy v organizaci tak, aby se papírových dokumentů objevovalo co nejméně.

- Pokud přece jen ke skenování dojde, snažte se textové dokumenty převádět do strojově čitelného formátu pomocí metody OCR.

Extrahování dat z webu neboli scrapování

Scrapování je technika extrahování dat z webových stránek. Moderní webové stránky jsou dynamicky generované, což znamená, že se obsah při každém načítání stránky generuje z dat v nějaké databázi. Zhruba do roku 2005 však většina webů měla statické stránky, kde byl obsah natvrdo a neměnně uložen. Možná máte na webu vaší instituce zajímavé tabulky, které někdo pracně naklikal a dnes nikde jinde neexistují. V takovém případě přichází ke slovu scrapování – ze strukturovaných

dat vepsaných do kódu stránky (obvykle v jazyce HTML) pomůže udělat zdrojovou tabulku.

Scrapování vyžaduje vedle dovednosti programování alespoň trochu zkušeností. Řešit to můžete pomocí nabídky projektů, jako je ScraporWiki.com, kde je k dispozici mnoho hotových skriptů. Ty pak můžete pro scrapování využít tak, jak jsou, případně si připlatit za oscrapování vybraných dat. Pokud jde o velký objem je to ekonomičtější varianta, než data přepisovat do tabulky ručně.

↳ Andrew Stott – Voláme hackery na pomoc:

Poradce britské vlády pro transparentnost a zakladatel portálu otevřených dat Opendata.gov.uk Andrew Stott má dobré zkušenosti s dobrovolnou pomocí ze strany vývojářské komunity. V Česku s takovým přístupem teprve experimentujeme. [První hackaton](#) [24] zorganizovaný veřejnou institucí pořádaly Institut plánování a rozvoje hlavního města Prahy s Fondem Otakara Motejla v červnu 2014. ←

Sensory, big data a „internet věcí“

Trendem desátých let [25] je „Internet of Things“ („internet věcí“, IOT). Jde zejména o levné senzory, které s maximální mírou detailu monitorují různé aspekty života ve městech. Relativně běžné jsou čítače dopravy nebo meteostanice. Trend bude posilovat, a to nejen proto, že řada firem odhalila, že senzory opentlená města generující obrovské objemy dat (tzv. big data) jsou skvělou příležitostí k výdělku. Ale také proto, že o životě ve městě máte konečně k dispozici reálná fakta.

Data primárně slouží k interní analýze, kterou dodavatel senzorů obvykle provádí sám. Než s ním ale podepíšete smlouvu, trvejte na tom, aby vám ke zveřejnění poskytoval i zdrojová data.

[24] Obvykle víkendové setkání vývojářské komunity, na kterém společně řeší nějaký (obecně prospěšný) úkol. Příkladem je Random Hacks of Kindness.

[25] Míněno 2010–2019.

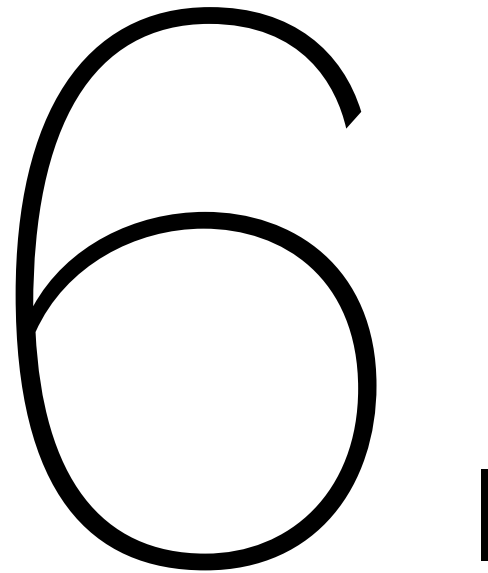
Jak datasety tvořit

Řadu datasetů v podobě tabulek úředníci sami tvoří. Právě tady často dochází k chybám – samotný autor se v rozsáhlé tabulce vyzná, už ji ale nepochopí kdokoli jiný, natož aby takový dataset propojoval s jiným, podobně nesrozumitelným.

Problém je možné řešit standardizací (struktura všech tabulek je dána nějakým vnitřním předpisem, případně ji kontroluje proškolený tým lidí) anebo vzděláváním. Dobrý základ pro získání představy, co je to datový model, dávají třeba lekce na webu DrawingByNumber.org nebo online [kurz Making Sense of Data](#) od společnosti Google.

↳ Co je datový model?

Datovým modelem rozumíme strukturu dat. Na příkladu tabulky v Excelu jde o první řádek – „nadpisy“ jednotlivých sloupců tabulky. To sice působí primitivně, ale možná právě proto se v tom dělá nejvíce chyb. V příloze uvádíme datové modely pro některé běžné datasety. ←



publikace dat

Drtivá většina dat – ať už jde o centrální, nebo lokální instituce – nemá to štěstí, aby se dostala k občanům ve srozumitelné podobě. Část z nich skončí na papírech kdesi v šuplíku, část se dostane na web instituce v podobě naskenovaných PDF dokumentů, které počítač nepřečte a pro člověka většinou znamenají manuální přepisování číslo po čísle.

Nejčastější problém tedy obvykle není v datech samotných (ačkoli špatně strukturovaná nebo nekompletní data opakované použití zcela znemožňují), ale v publikaci.

Manuální a automatizované zveřejňování

Pokud s otevřenými daty začínáte, nezapomínejte se otázkou automatického zveřejňování dat – může se stát, že o nabízený dataset nebude zájem a celá práce přijde vniveč. Začněte s několika vybranými datasety, které na web dejte ručně. Pokud zjistíte poptávku po pravidelné aktualizaci, má cenu se automatizací začít zabývat (více v kapitole Evaluace).

Manuální publikace (a její usnadnění)

Manuální publikací rozumíme vložení souboru na web. Jejím největším nebezpečím (zejména pokud data zveřejňuje víc lidí) jsou různé standardy a v důsledku praktická nepoužitelnost.

Máte-li ve své instituci CMS (Content Management System) pro správu obsahu webu nebo DMS (document management system) pro řízení oběhu dokumentů, bude pro vás publikace pravděpodobně snazší – samotné umístění dat na web bude probíhat stejně jako u jiných „příloh“, které na web běžně dáváte. Rozdíl bude pouze v popisu dat, tedy metadatech.

Manuální publikaci lze zjednodušit pomocí různých utilit („programků“). Příkladem je americká služba IFTTT.com (If This Then That – „když se stane tohle, udělej tamto“). IFTTT umožňuje nastavit například automatický post o publikaci dat na váš web poté, co soubor umístíte do Dropboxu (viz dále). Možností je mnoho, stačí trocha kreativity.

Automatická publikace

Několik firem nabízí platformy pro automatickou publikaci otevřených dat. Uvedme alespoň dvě: komplexní řešení americké firmy Socrata.com a evropský projekt Comsoode (Youropendata.eu). Automatizovat publikaci dat ale můžete i sami, stačí trocha pomoci programátora.

Publikace geodat

Geodata jsou specifická a je třeba je zmínit zvlášť. Nejlepší způsob publikace je pomocí otevřených webových služeb konsorcia Open Geospatial Consortium (OGC OWS), zejména pomocí Web Feature Service (WFS) a Web Coverage Service (WCS). Pomocí WFS publikujeme data vektorová (většina dat publikovaných státní správou), pomocí WCS publikujeme data rastrová (například letecké snímky). Existuje spousta open source serverových programů, které jsou nasazeny v celé řadě institucí. S jejich pomocí lze efektivně a jednoduše tyto služby nastavit a nabízet (například MapServer, GeoServer a další). Standardy OGC WFS a WCS jsou také používány jako tzv. stahovací služby podle směrnice INSPIRE.

Často se může stát, že jsou data publikována pomocí služby OGC Web Mapping Service (WMS). Je třeba zdůraznit, že v takovém případě se nejedná o otevřená data, protože výsledkem této služby je obrázek, tedy strojově nezpracovatelný formát bez původní informační hodnoty. Tuto službu lze s výhodou použít pro náhled dat, aby se uživatel mohl rozhodnout, která data chce/potřebuje, a udělal si základní představu o jejich podobě, ale nelze ji vydávat za publikační službu otevřených dat.

Kam data fyzicky umístit?

Úložiště dat by mělo splňovat alespoň následující požadavky:

- Dostatečná (ale ne přehnaná) disková kapacita
- Permanentní připojení k internetu
- Zálohování

Standardně se tyto požadavky řeší třemi způsoby, které uvádíme níže, v závěru oddílu je pak porovnávání v souhrnné tabulce.

Otevřená úložiště

Otevřená úložiště pravděpodobně znáte – patří mezi ně například server Ulož.to, proslulý zejména sdílením nelegálních kopií filmů (což se správci snaží omezit). V podstatě jde o to, že firma nabízí velkorosý diskový prostor výměnou za zobrazování reklamy, na které vydělává. Pro státní správu a samosprávu je to netypické řešení, nicméně není od věci ho zvažovat – splňuje totiž všechny výše uvedené podmínky, navíc je zcela zdarma. Nevýhody jsou dvě – jednak hrozí ztráta dobré pověsti (nemusí to vypadat důvěryhodně) a pak data nejsou pod kontrolou jejich vydavatele. To ale nemusí vadit – vzhledem k tomu, že jde o otevřená data, nemusíte se obávat jejich zneužití. Jediné vážné nebezpečí je výpadek, nebo dokonce zrušení služby – v takovém případě prostě data nahrajete jinam.

Doporučit můžeme dvě následující řešení:

● **Google Fusion Tables** (Dynamické tabulky Google)

Jde o nástroj zdarma nabízený firmou Google v rámci balíku cloudových kancelářských aplikací Google Drive. Jejich český název („dynamické tabulky“) není právě trefný a v komunitě uživatelů se stále užívá původní označení v angličtině.

K použití (nahrání dat) je třeba mít registrovaný Google účet, který je taktéž zdarma. Nástroj má bohatou dokumentaci, nahrávání dat a doplňování metadat je ale velmi intuitivní.

Mezi specifické výhody služby patří, že Google spravuje obrovské množství dat, nad kterými poskytuje velmi kvalitní vyhledávání. Nástroj tedy slouží také jako katalog dat; pravděpodobnost, že je někdo najde a opětovně využije, se tím výrazně zvyšuje. Uživatelům jsou navíc k dispozici analytické nástroje. Práce s daty probíhá přímo na serveru, který ale není váš, a proto vás jeho vytížení nemusí zajímat.

● **Cloud Storage**

Pod anglickým termínem se rozumí služby, které (obvykle zdarma nebo za symbolický poplatek) poskytují diskový prostor. Některé, jako je Ulož.to, se chovají jako obrovské (a málo tříděné) skladiště dat. Většina, jako je např. Úschovna.cz, Dropbox.com nebo Box.com, ale pracují s uživatelskými účty. V nich obvykle existuje jedna veřejná složka („Public“), do které kopírujete soubory k publikaci a získáváte jejich jedinečné webové adresy (URL). Ty pak vkládáte na web. Výhodou je, že data máte pohromadě v jedné složce ve vlastním počítači, v cloudu se jen zálohují.

Na závěr dodejme, že publikaci dat na otevřených úložištích doporučuje také Open Knowledge, jedna z výrazných světových autorit pro otevřená data [26].

Vlastní server

Větší instituce obvykle mívají k uložení dat vlastní server. Obvykle je třeba součinnost se správcem serveru a odpovědi na řadu bezpečnostních otázek.

Budete-li nad svými daty budovat API, pamatujte, že to znamená dlouhodobou zátěž v podobě provozu. Stanou-li se vaše data populárními (což si jistě přejete) a někdo na nich například postaví mobilní aplikaci, může nastat problém. Autor aplikace si totiž data nekopíruje k sobě, ale každé použití aplikace znamená online dotaz do dat na vašem webu (například při stažení mobilní aplikace s jízdními řády si do mobilu nestahujete celé jízdni řády, ale jen program, který potřebuje internetové připojení, aby se mohl kdykoli připojit k Idos.cz a dotaz na spoj položit za vás).

Popsaný způsob využití může znamenat mnoho požadavků (připojení) k vašemu serveru najednou a v důsledku jeho „spadnutí“ (to je princip i cílených útoků na servery, známých jako DoS útoky). Správci serverů tento problém řeší dvěma způsoby – buď zlepší hardwarové parametry (připojení k internetu, procesor), což samozřejmě znamená nemalé náklady, anebo nastaví nějaké omezení (například z jedné IP adresy je možné se k serveru připojit jen tisíckrát denně). Druhé řešení znamená

[26]

Relevantní námitkou proti umístění na veřejná cloudová úložiště je, že tím může utrpět důvěryhodnost takových dat. Pokud to ovšem řeší technické problémy, je taková cena patrně přijatelná.

omezení, a těžko pak data považovat za otevřená. Z toho důvodu (i když serverem disponujete) doporučujeme vybrat jinou variantu.

Placený cloud

Technicky totožné řešení, jako jsou otevřená úložiště, nabízejí placené cloudové služby. Za pravidelný poplatek získáte garanci a obvykle i lepší servis.

Způsoby zveřejňování

Soubor ke stažení

Nejčastější a nejjednodušší způsob zveřejňování. Soubory je možné pověsit jak na váš vlastní server, tak třeba do cloudu. Na web (do katalogu) vložíte prostý odkaz.

Neustále online – XML feed

XML feed je vhodný způsob zveřejňování, pokud

- vaše data bude jistě užívat někdo jiný
- jednotlivé položky se často mění
- data mají velký objem

Takové podmínky splňují například elektronické obchody: jejich zájmem je, aby o cenách jejich zboží věděly srovnávače cen, jako je Heureka.cz nebo Zboží.cz, nabídka a ceny se navíc často mění. Srovnávače cen samy pro takové webové nabídky vytvářejí standardy [27].

Nabídněte databázi co nejpřístupněji – přes API

API [28] (rozhraní pro programování aplikací) se stala v oblasti zveřejňování dat v zahraničí v poslední době velice populárními. Jde o rozhraní mezi databází a webem, které umožňuje programátorům klást dotazy a získávat odpovědi v podobě dat. První výhodou je,

že programátoři nemusejí stahovat celý balík dat, pouze část, relevantní pro položený dotaz. API navíc bývá aktualizováno v reálném čase, takže umožňuje okamžitě získávat nejnovější informace.

V Česku se API teprve zabydluje, přesto umožňuje přístup k řadě zajímavých dat: registrům státní správy (Registr územní identifikace, adres a nemovitostí – RÚIAN, Administrativní registr ekonomických subjektů – ARES, Český úřad zeměměřický a katastrální), bankám (Fio banka, Česká národní banka) i webovým komerčním službám (sReality, Kouzelná almara atd.). [Katalog českých API](#) udržuje Jan Javorek.

[Jak API vytvořit](#), radí řada webových fór a tutoriálů. Napsat ho můžete prakticky v jakémkoli jazyce – běžné je PHP, JSON nebo XML. Využít můžete i kolaborativních platforem. Speciálně pro API vznikl (původně český) produkt Apiary.io.

Na co si při tvorbě API dát pozor, upozorňuje [Joshua Bloch na Google Tech Talks](#) [29]. Vyplatí se též číst blogy, např. [API Evangelist.com](#) [30].

Webové služby pro geodata

Geodata je nejlépe publikovat pomocí webových služeb OGC WFS a WCS. Implementační pravidla pro směrnici INSPIRE dále doporučují nastavit pro stahování také formát ATOM.

Přidejte dokumentaci

Základní dokumentací jsou „data o datech“, tedy metadata – těm jsme se věnovali v podkapitole Přidejte data o datech. V mnoha případech

[27] Příklady: <http://sluzby.heureka.cz/napoveda/xml-feed/> nebo <http://napoveda.seznam.cz/cz/zbozi/napoveda-pro-internetove-obchody/specifikace-xml>.

[28] Oddíl je převzat z českého překladu Open Knowledge. The Open Data Handbook [online]. 2011 [cit. 2014-05-22]. Dostupné z: <http://opendatahandbook.org>.

[29] Bloch, Joshua. How To Design a Good API and Why it Matters. In: YouTube.com [online]. 2007 [cit. 2014-06-08]. Dostupné z: <https://www.youtube.com/watch?v=aAb7hSCtvGw>.

[30] Lane, Kin. API Evangelist [online]. [cit. 2014-06-08]. Dostupné z: <http://apievangelist.com>.

to bude také dokumentace postačující. Za ideální považujeme, pokud jsou strojově čitelná (více v kapitole Stažitelná a strukturovaná metadata).

Data ale mohou být složitá, jejich datový model nemusí být srozumitelný každému. V takovém případě je důležité zveřejnit i podrobnější dokumentaci, která navíc vysvětluje, jak jsou data získávána, nebo popisuje jejich datový model. Dobrým příkladem je [dokumentace k datům o hlasování v Poslanecké sněmovně Parlamentu České republiky](#).

Zařadte data do katalogu

Data se na nás sypou ze všech stran – norští badatelé přišli v roce 2013 s odhadem, že 90 % všech existujících dat vzniklo jen za poslední dva roky [31]. Dat je mnoho, klíčová je schopnost dohledat je. K tomu stále slouží katalogy.

↳ Katalogizujte!

Ať už s využitím dále zmíněné metodiky katalogizace, anebo vlastními cestami, v každém případě ale svá data katalogizujte. ⇐

Národní katalog

Česká vláda slíbila vytvořit národní katalog otevřených dat již v roce 2012 v rámci svého Akčního plánu Partnerství pro otevřené vládnutí (Open Government Partnership, OGP). Pro Úřad vlády ČR pak vznikla [Koncepce katalogizace otevřených dat](#), z níž v této knize vycházíme. Přestože právě u koncepce zatím práce na katalogu skončily, předpokládáme, že se ledy rozhýbou a katalog [na portálu státní správy](#) vznikne.

Oficiální katalog [32] otevřených dat vytváří čím dál více zemí. Za vzorový bývá vydáván například [katalog britský](#), [americký](#) nebo [švýcarský](#).

Minikatalogy

Považujeme za dobrý nápad, pokud každá instituce připraví vlastní katalog dat. Jednak je to vstřícný krok vůči uživatelům vašeho webu, jednak takový krok pomáhá dohledatelnosti.

Pod minikatalogem si netřeba představovat nic složitějšího – může

jit o obyčejnou tabulku na webových stránkách. Český telekomunikační úřad vytvořil [katalog](#), v němž je možné i filtrovat.

Geodata

Česká státní správa a samospráva buduje geografické katalogy převážně díky směrnici INSPIRE. [Národní geoportál INSPIRE](#) sbírá metadata ze serverů podřízených organizací a umožňuje tak rychlé vyhledávání nad těmito metadatovými záznamy.

Služby fungující jako katalogy

Primárním účelem následujících (a mnoha dalších) služeb není funkce katalogu, ale porovnávání a hodnocení otevřených dat. Přesto není k zahození je znát, pro základní přehled o tom, jaká data (zvláště státní správy) existují.

The Open Data Certificate (Certifikát otevřených dat)

Novinka The Open Data Institute si klade za cíl zvyšovat kvalitu dat, mimoděk ale vzniká jejich [katalog](#). O certifikátu píšeme v kapitole Kvalita dat.

Open Data Index (Žebříček otevřených dat)

[Index organizace Open Knowledge](#) se zaměřuje jen na deset vybraných vládních datasetů.

[31] Dragland, Åse. Big Data – for Better or Worse. [online]. [cit. 2014-05-22]. Dostupné z: <http://www.sintef.no/home/Press-Room/Research-News/Big-Data-for-better-or-worse>.

[32] Početně samozřejmě převažují katalogy neoficiální; katalogizační funkci mají i projekty jako Open Data Index nebo Open Data Certificate. Garance státu je však důležitá: spolu s daty získáváte jistotu, že za nimi stojí autorita s vahou státní moci.

7

**kvalita
dat**

V předcházejících kapitolách bylo cílem zveřejnit a otevřít data. Pod heslem „raději hned a hůř než dokonale a nikdy“ jsme se snažili poslat data do světa, k jejich uživatelům. Teď je na čase zaměřit se na ono „dokonale“. (Pokud vám některé koncepty budou povědomé – ano, už jsme je zmiňovali v předchozím textu. Tentokrát ovšem jdeme víc do hloubky.)

Pětihvězdičkové schéma

Internetový vizionář a vynálezce webu Tim Berners-Lee sestavil pro hodnocení otevřenosti dat pětibodovou škálu, aby zdůraznil, co je při zveřejňování skutečně důležité. Z této se i díky osobnosti jejího tvůrce stal de facto standard. Stejně jako u hotelů platí, že čím víc hvězdiček, tím líp:

*

Publikujte data online, dovoďte je opětovně použít

I kdyby mělo jít o naskenované dokumenty, publikujte je alespoň online. Bude-li o ně zájem, uživatelé si je sami do strojově čitelné podoby převedou. Například Státní okresní archiv v Třeboni byl před lety zavalen amatérskými genealogy, kteří toužili hledat v matričních knihách. Archiv tedy knihy naskenoval a vytvořil pro pohodlné užívání vlastní prohlížeč. Bylo by zbytečné všechna data přepisovat (většina ze zápisů je psána švabachem, OCR by proto nebylo možné efektivně použít), stačilo, že jsou data online. Archiv tak šetří prostředky sobě (nemusí obsluhovat zástup zájemců) i uživatelům (kteří nemusí dojíždět do jižních Čech).

K tématu licencí se podrobně vyjadřujeme v podkapitole Publikujte bez právních omezení.

**

Strojově čitelná data

Dvuhvězdičková data nejsou jen online, ale též strojově čitelná. Tedy srozumitelná nejen pro člověka, ale i pro počítač. V praxi to znamená, že data musí být v podobě znaků, nikoli obrázků. Typickým příkladem dvuhvězdičkových dat jsou excelové tabulky, v podstatě z každé databáze ale můžete data exportovat za dvě hvězdičky.

(Relevantní poznámku k tomuto bodu má Zákon č. 106/1999 O svobodném přístupu k informacím. Judikatura říká, že úřad musí informaci poskytnout v takové podobě, v jaké ji vytvořil a uchovává, typicky tedy jako dvuhvězdičkový XLS nebo tříhvězdičkový CSV. Praxe, kdy úřad tabulku vytiskne, oskenuje a nabídne jako jednohvězdičkové PDF, je tedy nezákonná.)

Otevřené formáty

Tříhvězdičková data považujeme za standard pro data veřejné správy/samosprávy. Otevřenými formáty pak myslíme požadavek na to, aby dataset mohl použít uživatel s běžným softwarem. Typickým příkladem je rozdíl mezi formáty XLS a XLSX (přestože oba pocházejí z dílny Microsoftu). První je uzavřený, zatímco s druhým je možné pracovat i pomocí alternativních programů, jako je LibreOffice nebo OpenOffice. Při publikaci ale častěji narazíte na problémy s formáty geodat: komerčně dodávaný software obvykle pracuje s vlastními formáty, nečitelnými kdekoli jinde.

Univerzální identifikátory

Čtyřhvězdičková data chápeme v českém prostředí jako nadstandard. Klíčovým pojmem je „univerzální identifikátor“ (URI) – zní to tajemně, ale nejde o nic jiného než přiřazení vlastnosti každému sloupci tabulky. To znamená, že kromě člověka pochopí obsah tabulky bez nějaké zásadní dopomoci i počítač. K tomu je potřeba pojmenovat sloupce podle určité konvence. Technologické řešení takové strojově srozumitelné struktury vyžaduje buď důslednost (standardizaci) anebo kvalifikovanou ruční práci (RDF skeleton).



URI skrze důslednost: standardizace

Odbor Hlavního architekta eGovernmentu na ministerstvu vnitra zavedl vyhláškou č. 469/2006 Sb. Informační systém o datových prvcích, [metodika](#) k ní je k dispozici na webu ministerstva. Ta zjednodušeně říká, jak má záhlaví každé tabulky (datový model) jednotně vypadat. Při porovnání stejných datasetů různými institucemi se tak mělo zamezit problému

neporovnatelnostinebonepřenositelnostidat.Takovýprocesbýváoznačovánjako standardizace,jehoobecnýproblémvšakje,ževyžadujedůslednost.Informační systém o datových prvcích tedy oficiálně funguje a je třeba ho využívat.

● URI skrze manuální práci: RDF skeleton

Nejprogressivnější technologií pro zavádění URI je Resource Description Framework (RDF). Filosofie RDF je založena na přiřazování vlastností k jednotlivým objektům, které mají jedinečné URI.

Jako příklad poslouží třeba otec teorie informace Claude Shannon. Toho můžeme identifikovat jedinečným URI třeba přes jeho profil na Wikipedii s jedinečnou webovou adresou: http://cs.wikipedia.org/wiki/Claude_Shannon. Tím máme definován objekt, kterému můžeme přiřazovat vlastnosti. Vznikají tedy trojice subjekt – vztah – objekt. Například

http://cs.wikipedia.org/wiki/Claude_Shannon JE MATEMATIK

nebo

http://cs.wikipedia.org/wiki/Claude_Shannon NAPSAL
http://en.wikipedia.org/wiki/A_Mathematical_Theory_of_Communication.

Přiřazování vlastností je možné provést psaním kódu, k zefektivnění ale existují programy, které práci omezují na „naklikání“ zmíněných trojic. Oblíbeným řešením je rozšíření softwaru OpenRefine.org o freewarový doplněk LODRefine.

● Jiné technologie: mapy témat (Topic Maps)

Mapy námětů [33] jsou pouhou technologickou alternativou k RDF. Smysl je stejný – totiž dát dílčím informacím smysl doplněním kontextu. Kontext a propojení jsou reprezentovány jakousi „myšlenkovou mapou“. Koncept je svou kvalitou s RDF srovnatelný, v jeho konkurenci se nicméně neprosadil. Přesto na něm funguje řada úspěšných projektů, jako jsou Zakonyprolidi.cz [34].

↳ Sémantika, RDFa, mikroformáty, web 3.0...

V oboru kolem čtyřhvězdičkových dat se používá řada odborných pojmů. Vysvětlujeme je v glosáři. Mírný zmatek způsobuje to, že jde častokrát o synonyma, která se liší jen technologickým řešením. Jakmile ale jednou pochopíte princip, měli byste bez problémů rozumět názvosloví. ←

↳ Sémantický web:

Pro orientaci v tématu je vhodné zmínit i pojem sémantický web. Je to systém, který některým termínům v textu webových stránek přiřazuje význam.

Ukažme si to na příkladu webu této publikace. Návštěvník webu vidí toto: Jak otevírat data, jehož autorem je Jakub Mráček, podléhá licenci Creative Commons Uvedte autora-[Zachovejte licenci 4.0 Mezinárodní](#)

V kódu stránky je ale o něco víc:

```
<span xmlns:dct=„http://purl.org/dc/terms/“ property=„dct:title“>
Jak otevírat data</span>, jehož autorem je <span xmlns:cc=
„http://creativecommons.org/ns#“ property=„cc:attributionName“>
Jakub Mráček</span>, podléhá licenci <a rel=„license“ href=
„http://creativecommons.org/licenses/by-sa/4.0/“>Creative Commons
Uvedte autora-Zachovejte licenci 4.0 Mezinárodní</a>.
```

Informace v hranatých závorkách slouží vyhledávači, který stránku prochází. Ten díky nim ví, že „Jak otevírat data“ je název díla, „Jakub Mráček“ je jméno a „Creative Commons“ licence.

Z výše uvedeného je zřejmé, že na rozdíl od tříhvězdičkových dat znamená zavádění sémantických dat skokově více práce (zároveň si ovšem připravujete půdu k propojování dat). Dříve, než se do jejich budování pustíte, nezapomeňte provést analýzu nákladů a výnosů, případně se poradit s odborníky. ←

[33] Český termín se prakticky nepoužívá.

[34] Systém, na kterém projekt běží, dostal jméno ATOM: <http://demo.atom2.cz/form/default.aspx>.

Propojená data

O propojených datech (Linked Data) hovoříme v kapitole Kam dál: Propojená data. V kontextu pětihvězdičkového schématu si jen řekněme, co je oproti předchozí úrovni třeba udělat navíc.

Propojení dat spočívá v tom, že minimálně u dvou datasetů (pomocí kódu) určíte, která data jsou společná. Typickým příkladem jsou statistické údaje o jednotlivých zemích, které publikuje Světová banka. V jedné tabulce bude název země a počet obyvatel, ve druhé název země a HDP. Jak dostat oba údaje do jednoho datasetu? Řeknete, že názvy zemí v jedné tabulce jsou ekvivalentní názvům zemí v té druhé. Pomocí dotazovacího jazyka **SPARQL** se pak můžete dotazovat na počet obyvatel i HDP najednou, přestože každý údaj se nachází v jiné tabulce.

OL RE OF URI LD
http://data...+http://data...

OL RE OF URI
http://data...

OL RE OF
CSV

**
OL RE
XLS

*
OL
PDF

↳ Snažte se o ***

Prakticky použitelná jsou data, která dosahují alespoň na tři hvězdičky. Kompletní popis včetně řady příkladů najdete na webu 5stardata.info. ↩

↳ Open Data Ready:

Na Slovensku nedávno vznikla metodika pro nové státní IT projekty. Ty musejí být stavěny tak, aby jejich datové výstupy byly „Open Data Ready“. Pro Česko je situace zajímavá především způsobem, jak téma zakotvit legislativně. Principiálně se ale pohybujeme mezi stejnými požadavky, jaké uvádíme výše. ↩

Formáty dat v detailu [35]

Formáty dat jsme v předchozím textu spíš prolétli a neférově je pouze vyjmenovali bez podrobnějšího vysvětlení. Tady to napravujeme.

JSON

JSON je formát, který je snadno čitelný pro jakýkoli programovací jazyk, a potažmo tedy snadněji zpracovatelný pro počítače než například XML. Oproti němu má jedinou nevýhodu, obvykle nemá zaručenou formální strukturu [36].

XML

XML je široce používaný formát pro výměnu dat, jelikož umožňuje zachovat neporušené jak informace, tak i vzhled souboru. Tvůrce tedy může data bez obav naformátovat, aniž by došlo ke ztrátě původní informace.

RDF

RDF je formát doporučený konsorciem W3C, které stojí za mnoha webovými standardy. Umožňuje zobrazovat data ve formě vhodné pro kombinování informací z více zdrojů. RDF data mohou být uchovávána například v podobě XML nebo JSON. RDF formát není doposud příliš rozšířený, nicméně již došlo k jeho použití v rámci projektů

[35] Pasáž o formátech je český překlad The Open Data Handbook – Open Knowledge. The Open Data Handbook [online]. 2011 [cit. 2014-05-22]. Dostupné z: <http://opendatahandbook.org>.

[36] JSON schema umožňuje validovat samozřejmě i JSON.

propojených otevřených dat ve Španělsku a Velké Británii. Vynálezce webu Tim Berners-Lee považuje RDF za jeden z cílů projektů otevřených dat.

Tabulkový procesor

Mnoho úřadů má své informace uchovány v tabulkových procesorech (například MS Excel). Informace mohou být použity okamžitě za použití popisku, co daná buňka znamená.

Nevýhodou je, že některá makra nebo formuláře mohou být nefunkční v jiných než původních programech. Je proto dobré přiložit k tabulkovému procesoru i dokumentaci toho, co soubor obsahuje. Další překážkou může být uzavřenost formátu pro proprietární software.

CSV (Comma Separated Values)

CSV soubory jsou skvělé v tom, že mají pevně daný tvar, a jsou proto vhodné pro přenášení velkých balíků informací se stejnou strukturou. Formát informací je však natolik jednoduchý, že je prakticky nepoužitelný bez přiloženého popisu – bez hlavičky je v zásadě nemožné přijít na to, co obsahují jednotlivé sloupce. Z tohoto důvodu je důležité k CSV souborům vždy přikládat příslušnou dokumentaci.

Nejnutnější je dbát na přesnost – i sebemenší chyba v jedné buňce může způsobit nepoužitelnost zbylých dat bez možnosti dohledat jejich původní význam.

„Wordovský“ dokument

Klasické dokumenty jako DOC, ODF nebo OOXML jsou dobré na zobrazování relativně stálých informací – třeba mailing listu. Zveřejňovat tyto formáty bývá velice levné, protože jsou vytvořeny v určitém programu a pak vyvěšeny na internet (neprochází konverzí). Nevýhodou však je to, že dokumenty nedrží přesnou strukturu, což v praxi znamená, že je těžké je strojově číst. Z tohoto důvodu je nutné používat standardizované šablony, které umožní další automatické zpracování.

Raději ale pro zpřístupnění strukturovaných dat použijte jiný formát.

Čistý text

Čisté textové dokumenty (TXT) jsou velice dobře strojově čitelné. Problém však je, že obvykle neobsahují informace o struktuře přímo v dokumentu, což nutí programátory vytvořit tzv. parser, který dokáže data interpretovat.

Určité problémy může také způsobovat použití textového formátu v různých operačních systémech – typickým problémem je fakt, že MS Windows, Mac OS X a jiné Unix systémy mají vlastní způsoby, jak říci počítači, že dosáhly konce dokumentu.

Naskenované obrázky

Jedná se pravděpodobně o nejméně vhodný formát pro otevřená data, ale TIFF a JPEG-2000 jsou alespoň schopny nést odkaz k dokumentaci, která vysvětlí, co je na naskenovaném obrázku. Bývají používány pro zobrazování informací, jež nemají elektronický původ (například staré církevní záznamy), u čehož by i zůstalo, kdyby nedošlo ke skenování.

Geoformáty

Potřebujeme-li publikovat data, která mají nějakou prostorovou informaci, je nejlepší sáhnout po některém z rozšířených a uznávaných formátů pro prostorová data. Ve světě geografických informačních systémů existuje standardizační organizace [OpenGeospatial Consortium](#) (OGC), která se mimo jiné stará o standardy Geography Markup Language (GML) a Keyhole Markup Language (KML). Druhý zmiňovaný dosáhl popularity díky firmě Google, která je jeho původním autorem a prosazuje ho ve svých produktech. Oba formáty jsou postaveny na XML.

Vedle de iure standardů OGC existuje řada otevřených de facto standardů – rozšířených formátů. Z těch starších, a proto široce podporovaných jde například o ESRI Shapefile (SHP) nebo novější GeoJSON (JSON s geosložkou).

Máme-li geodata například v tabulce, je vhodné publikovat je ve formátu CSV a k tomuto souboru ještě vytvořit metadatový soubor VRT, který v sobě nese informace o souřadnicovém systému a o tom, ve kterém sloupečku se nachází jaká souřadnice. Popis, jak se tvoří, najdete [zde](#).

Nezapomeňte na to, že geodata jsou často v různých souřadnicových systémech. V jiném systému měří zařízení GPS, v jiném jsou například mapy katastru nemovitostí. Pokud známe souřadnicový systém, je vhodné informaci o něm k datům přiložit (většina formátů ji obsahuje automaticky, tabulková data nikdy). Pro rastrová data (obrázky, letecké snímky) slouží například formát GeoTIFF.

Individuálně vytvořené formáty

Některé specifické systémy mají vlastní datové formáty, ve kterých ukládají nebo exportují informace. Někdy stačí informace zpřístupňovat v těchto formátech, především za situace, kdy se předpokládá využití v programech podobných těm, ve kterých byl dokument vytvořen. K dokumentům by měl být vždy přikládán odkaz na informace o příslušném formátu (například na stránky výrobce). Obecně však platí, že otevřená data by neměla být zobrazována v individuálně vytvořených formátech.

HTML

V dnešní době jsou data často zobrazována v HTML. Tento formát je obzvláště užitečný, pokud se data často nemění a mají přiměřený objem. Pochopitelně je lepší zpřístupňovat je ve formátu jednodušším pro stahování, ale vzhledem k nákladům potřebným pro vytvoření HTML se jedná o dobrou cestu, jak začít se zveřejňováním.

Pokud používáte HTML, je důležité zapisovat informace do tabulek, kterým přidělíte ID, což umožňuje snadnější vyhledávání a manipulaci s daty. Yahoo vytvořilo [nástroj](#), který umí získat strukturovanou informaci z internetové stránky, přičemž jeho účinnost se zvyšuje, pokud jsou data řádně označena.

Technická omezení na vlastním serveru

Pokud publikujete data na vlastním serveru, pravděpodobně narazíte v debatách se správcem sítě na dva technické problémy: omezení počtu přístupů z jedné adresy – kvůli bezpečnosti – a trvanlivost odkazů, na nichž budou data viset.

Omezujte počet přístupů z jedné adresy co nejméně

Správci serverů často jako prevenci před DoS útoky omezují počet denních přístupů pro jednu IP adresu (tedy z jednoho počítače). Takové omezení je rozumné, protože zaručuje stálou dostupnost pro všechny uživatele. Zároveň ale omezuje scrapování nebo obecně hromadný přístup.

Doporučujeme proto několik zásad, které splní očekávání v oblasti bezpečnosti a dostupnosti, a přitom nebudou většinou překážet:

- Limitní počet přístupů pro jednu IP adresu nastavte co nejvyšší.
- V dokumentaci k datasetům toto omezení zmiňte a uveďte kontakt na správce serveru pro případ, že by někdo potřeboval větší počet přístupů.
- Evidujte IP adresy, z nichž dotazy přicházejí. Po čase vyhodnoťte, jaké jsou reálné potřeby uživatelů (může se stát, že riziko je pouze teoretické a omezovat přístupnost kvůli virtuálním problémům je zbytečné).
- Umožněte-li dump (export) celé databáze, počet dotazů z jedné adresy se sníží [37].

Neměňte adresy datasetů [38]

Jaká je šance, že vám pošta doručí noviny, když se přestěhujete a změnu nikde neohlásíte? Pravděpodobně nízká. Stejně je to s daty – pokud je vy (anebo váš systém) ustavičně stěhujete, přidáváte další překážku pro uživatele navíc. Řešení je v principu snadné a má dvě podmínky:

- Každý dataset bude mít unikátní adresu (URL).
- Žádná z těchto adres se nebude měnit [39].

[37] Pak ovšem budete řešit datový tok.

[38] Perzistentní URL, případně perzistentní URI považujeme za synonyma (URI je obecnější termín, vyjadřující jednoznačný identifikátor; URL je jeho reprezentace). Více o této problematice v doporučeních W3C (<http://www.w3.org/Provider/Style/URI.html>), <http://www.w3.org/Consortium/Persistence.html>), případně http://en.wikipedia.org/wiki/Persistent_uniform_resource_locator.

Dobrá permanentní adresa splňuje několik pravidel, která autoři doporučení pro úřady EU [40] shrnují do 10 pravidel pro permanentní URI:

Co dělat

- 1 Držte se šablony.
`http://{doména}/{typ-zdroje}/{koncept}/{identifikátor}`
- 2 Využívejte zavedené identifikátory.
`http://posta.cz/id/psc/68001`
- 3 Odkazujte se na alternativní formáty.
`<link rel=„alternate“ href=„dokument.rdf“>`
- 4 Zaveďte URI pro fyzické objekty.
`http://napriklad.cz/id/jan_novak`
- 5 Použijte stabilní a nezávislou doménu.
Permanentní URI by neměla být závislá na vydavateli dat.

Co nedělat

- 6 Neuvádějte vydavatele dat.
`http://vzdelavani.cz/msmt/id/skola/12345`
- 7 Neuvádějte číslo verze.
`http://vzdelavani.cz/doc/skola/v1/123456`
- 8 Nepoužívejte automatické vzestupné číslování.
`http://vzdelavani.cz/id/skola/123456`
`http://vzdelavani.cz/id/skola/123457`
- 9 Nepoužívejte query parametry.
`http://vzdelavani.cz/skola?id=123456`
- 10 Nepoužívejte koncovky souborů.
`http://vzdelavani.cz/doc/skola/123456.csv`

Tuzemským příkladem je skladba URL na portálu Volby.cz: z adresy `http://www.volby.cz/pls/ps2010/ps411?xjazyk=CZ&xobec=544281&xpm=0` lze docela dobře vydedukovat, jak je adresa stavěná, a analogicky konstruovat jinou.

Strojově čitelná a strukturovaná metadata

Podstatnou součástí každých dat je pohádka o jejich původu, struktuře a dalších vlastnostech – metadata. Je proto třeba přistupovat k nim podobně jako k samotným datům. Měla by být strojově čitelná, strukturovaná a snadno stažitelná (třeba jako XML nebo CSV), ideálně pro každý dataset zvlášť i pro všechna vámi nabízená data najednou.

Ontologie a jejich využití

Ontologie [41] jsou způsob, jak univerzálně popsat data. Příkladem je například Schema.org [42]. Některé obory si vyvinuly vlastní rozsáhlé ontologie, jako je ontologie pro popis dat v humanitární oblasti.

Evropská unie prosazuje vlastní [slovník DCAT-AP](#), který je odvozen od doporučení [DCAT](#) (Data Catalog Vocabulary) konsorcia World Wide Web (W3C). Jak název napovídá, slovník je určen pro popis datasetů v katalogu (definuje tedy podobu metadat). Jeho doplňkem je tezaurus [EuroVoc](#) [43].

Standardizace standardů

Neodpustíme si drobnou jízlivost: žádný standard není dokonalý, možná právě proto má tolik lidí potřebu tvořit standardy stále nové a (pochopitelně) dokonalejší. Bohužel tak standardů spíše přibývá a na konferencích se žertuje o tom, že bude třeba vytvořit standard pro standardy. Jízlivost tedy nahradíme apelem: Respektujte doporučené standardy, jinak se z bludného kruhu nikdy nevygotáme!

[39] Adresa nebude generována dynamicky.

[40] D7.1.3 – Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC [online]. [cit. 2014-05-24]. Dostupné z: <https://joinup.ec.europa.eu/sites/default/files/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf>.

[41] Ontologie v infromatickém kontextu znamená explicitní formalizovaný popis (například záznamu v databázi). Obsahuje glosář (definici pojmů) a tezaurus (vztahy mezi jednotlivými pojmy).





[42] Schema.org má pro Dataset metadata dokonce standardizovaná: <http://schema.org/Dataset>.

[43] DCAT-AP používají mimo jiné Evropský parlament, Úřad pro publikace, parlamenty členských států EU a jejich regionů, správní orgány členských států a soukromí uživatelé z členských i nečlenských zemí.

Certifikát The Open Data Institute

Certifikát The Open Data Institute je tím nejlepším pomocníkem při zvyšování kvality dat, který jsme poznali. Respektovaná instituce nabízí (zcela zdarma) certifikát pro jakýkoli dataset, výměnou žádá poctivé vyplnění komplexního dotazníku.

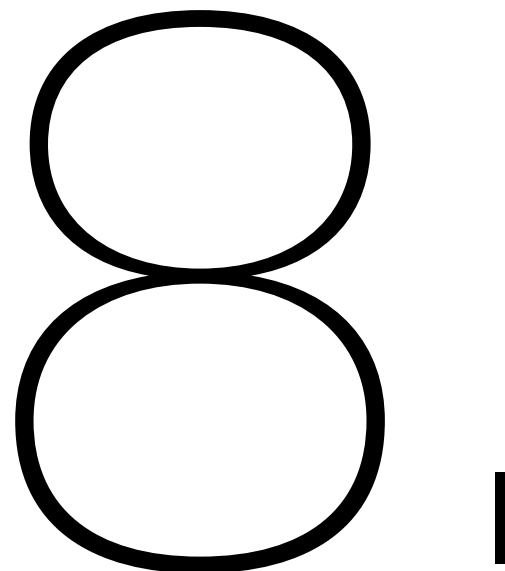
Certifikát má čtyři úrovně:

- | | | |
|---|-------------------|---|
| 1 | Surová data (Raw) |  |
| 2 | Pilot |  |
| 3 | Standard |  |
| 4 | Expert |  |

V každém okamžiku (už během vyplňování) vidíte, co je třeba doplnit, abyste dosáhli na vyšší úroveň. Nenápadnou gamifikací tak autoři dosahují zvyšování kvality dat.

Vygenerovaný certifikát je možné vložit i na svůj web, systém navíc funguje jako datový katalog.

Aplikace pro jeho získání je k dispozici [zde](#). Díky spolupráci Fondu Otakara Motejla s The Open Data Institute lze v průběhu roku 2014 očekávat i českou jazykovou mutaci.



Neaktualizovaný web je pro jeho správce špatnou vizitkou. Podobně i neudržovaná data a jejich katalog mohou značně zhoršit vaši reputaci. Proto je už při plánování procesu publikování otevřených dat třeba myslet na údržbu dat.

Co všechno údržba zahrnuje

- **Zajištění funkčnosti webu**

Hlídejte technickou stránku věci, přístupnost. Předvídejte vlny přístupů a technicky se na ně připravte [45].
- **Opravy chyb**

V datasetech i v metadatech bývají chyby. Sbírejte na takové chyby upozornění, s uživateli o nich otevřeně komunikujte.
- **Zvyšování kvality již publikovaných datasetů**

Datasety, o které je zájem, má smysl zkvalitňovat. Věnujeme se tomu v předchozí kapitole Kvalita dat.
- **Aktualizace datasetů**

Data mohou zastarávat. Nezapomeňte v metadatech uvést, zda se jedná o jednorázové zveřejnění, anebo stav k danému datu. Je také vhodné uvést, v jakém intervalu budou data aktualizována
- **Doplňování nových datasetů**

V neposlední řadě uživatelé přivítají, budete-li vedle údržby „starých“ datasetů přidávat i nové.

Náročnost a automatizace procesů

Každá práce něco stojí, a tak je důležité plánovat, kolik jí údržbě dat můžete a chcete věnovat.

Záleží na tom, jaká data, jakým systémem a v jakém objemu publikujete. Z materiálu Koncepce katalogizace otevřených dat pro Úřad vlády ČR vyplynulo, že národní katalog by měl spravovat jeden člověk na plný úvazek. Tiskový mluvčí České obchodní inspekce věnuje údržbě dat každý týden 15 minut. Zátěž tedy nemusí být vysoká, pokud máte proces publikace i opravy chyb dobře nastavený a (pokud možno) i zautomatizovaný.

Zapojte komunitu

Přestože je to z pohledu státního úředníka často nemyslitelné, ve Velké Británii mají výborné zkušenosti se zapojováním online komunity. Nejvýše postavený britský úředník přes otevřená data, Andrew Stott, poslal po Twitteru výzvu hackerům a vývojářům, aby pomohli opravit chyby v datech o všech autobusových zastávkách ve Spojeném království. Světe div se, věc se podařila. Pokud se k takovému kroku odhodláte, můžete buď kontaktovat českou komunitu kolem otevřených dat na mailing listu [Open Knowledge Foundation](#) nebo například programátorskou „klubovnu“ [GDG Garage](#) či třeba organizátory hackatonů, jako jsou Random Hacks of Kindness či DevCamp.

[45]

To platí jen v některých případech: například server Volby.cz, provozovaný Českým statistickým úřadem, je zatěžován během voleb a krátce po nich. V takovém období je vhodné zvýšit kapacitu připojení nebo připravit záložní servery.

9 |

**evaluace:
data
generují data
jak je čist?**

Digitální služby mají jednu sympatickou vlastnost – generují data a pomáhají tak vylepšovat samy sebe. Věnujte průběžnému vyhodnocování pozornost, ušetříte si tak spoustu zbytečné práce.

Co uživatelé doopravdy chtějí?

Na mnoha místech v procesu otevírání dat pracujeme s heuristikou: ať už na základě doporučení (třeba této knihy), zkušeností odjinud nebo prosté intuice více či méně kvalifikovaně hádáme, o co mají naši potenciální uživatelé zájem.

Sekce pro digitální služby na britském úřadu vlády, která koordinuje online služby celé státní správy Spojeného království, pracuje agilním přístupem. Své služby (včetně otevírání dat) postupně testuje a zjišťuje, co uživatelé doopravdy chtějí (nikoli, co my si myslíme, že chtějí). V mnoha článcích na jejich [blogu](#) narážíme na koncept „*Desire Paths*“ (vytoužené cesty) – proces „komunikace“ s uživatelem přenesený z urbanismu. Design cest v parcích nebo na náměstích se nerodí „na rýsovacím prkně“ architekta, ale přímo na místě. Plocha se osází jen trávou a čeká se, kudy lidé vyhodí cestičky. Ty se pak „potvrdí“ vydlážděním a obrubníky.

V designu služeb existuje celá řada podobných postupů, které spojuje snaha uživatelům skutečně porozumět a stavět služby podle skutečných požadavků. Zájemcům o tuto problematiku doporučujeme knihu *Skvělé služby* ([Designslužeb.cz](#)) od autorského kolektivu kolem Adama Hazdry.

Metrika: Návštěvnost a počet stažení

Klíčovou metrikou ukazující zájem o vaše data je návštěvnost, resp. počet stažení. U každého datasetu ji sledujte a vyhodnocujte.

↳ Google Analytics:

Google nabízí zcela zdarma [analytický nástroj](#), který vám pomůže sledovat jak jednotlivé metriky – třeba návštěvnost v čase – tak odpovídat na složitější otázky. Implementace nástroje není složitá, využít však můžete řadu kurzů, třeba přímo od Googlu. ◀

Co víte o svém uživateli?

Moderní analytické nástroje, jako je Google Analytics, dokážou o uživateli vašich dat říct mnohem víc. Příklady otázek, které si můžete pokládat, jsou například:

- Přistupují uživatelé k datům z vašeho webu, anebo přesměrováním z externího odkazu? Pokud z externího odkazu, zaměřte se na spolupráci se stránkami odkazujícího partnera.
- Jaký formát dat stahují u jednoho datasetu uživatelé nejčastěji (nabízíte-li jich víc)? Pokud jen jeden, ušetřete příště čas na exporty do dalších formátů.
- Jaký je poměr přístupů z mobilních zařízení? Pakliže vysoký, je váš web pro mobilní zařízení přizpůsoben?
- Jak často uživatelé hledají frázi „otevřená data“? Pokud často, pravděpodobně je sekce „otevřená data“ na vašem webu pro většinu uživatelů obtížně dohledatelná.
- Kolik stránek uživatelé proklikají, než se dostanou k datasetu? Pokud hodně, změňte strukturu nebo navigaci.

Co se s daty děje dál?

Součástí evaluace je také přehled o tom, k čemu se vaše data využívají. Výsledkem mohou být lepší argumenty dovnitř instituce („Naše data využily tři celostátní deníky k zajímavé analýze“ nebo „Soukromá firma postavila na datech aplikaci, v níž nám dělá reklamu“), ale také zlepšení vašich služeb.

Není od věci požádat uživatele, kteří si data stáhnou, o kontakt nebo o zpětnou vazbu. Zároveň ale není možné kontaktem přístup k datům podmiňovat.

Jak využití dat podporovat (a zpětně z toho profitovat)?

Fond Otakara Motejla pořádá pro vývojáře soutěž **Společně otevíráme data** o nejlepší aplikaci postavenou na otevřených datech. Stejně jako Nadace Vodafone v prvním ročníku má kdokoli možnost stát se partnerem soutěže a podpořit využití vlastních datasetů (pokud jsou otevřené).

Existující aplikace nad otevřenými daty pak můžete přihlásit do řady soutěží, mj. nejprestižnějšího ocenění českých internetových projektů **Křišťálová lupa**. Od roku 2012 zde existuje kategorie Veřejně prospěšný projekt/aplikace.

10.

kam dál: propojená data

Malé překvapení na závěr: doteď jsme obvykle mluvili o otevřených datech jako maximu, kterého bychom chtěli dosáhnout. V tuhle chvíli jsou pro veřejnou správu skutečně aktuální. Už teď ale víme, co přijde po nich. Vizionář Tim Berners-Lee před několika lety popsal budoucnost otevřených dat v pětihvězdičkovém schématu. To, co jsme popisovali v této příručce, dosahuje na dvě nebo tři hvězdičky. Čtvrtá a pátá hvězda znamenají otevřená propojená data [46] – tedy propojení databází různých institucí na základě společných pravidel pro jejich strukturu.

Idea je velmi prostá: pokud na jednom místě o Timovi Berners-Leeovi napíšete, že je vynálezcem webu, a na jiném, že pracuje ve W3C (a pokud jsou obě místa propojená), pak mohou oba dva přívlastky mít na jednom místě.

Kde a jak propojená data fungují?

Společným příznakem následujících (a mnoha jiných) příkladů je obrovský objem strukturovaných dat – tak obrovský, že obvykle nemá konkurenci. I to je jednou z motivací pro propojování.

↳ Obnovitelné zdroje na jednom místě:

Reegle.info je databáze dat o obnovitelných zdrojích. Data přitom nejsou původní, jde o propojení [47] deseti velkých poskytovatelů dat, jako je Světová banka nebo Mezinárodní energetická agentura.

Informace o lécích v jednoduché databázi:

Léková encyklopedie propojuje data z pěti různých zdrojů. Výsledkem je unikátní aplikace umožňující najít k léku jeho varianty nebo kontraindikace. Projekt byl oceněn v rámci soutěže **Společně otevíráme data** v roce 2013.

Cesta Wikipedie z datové džungle k Wikidata a DBpedii:

Když v roce 2001 zakládal Jimmy Wales největší encyklopedii na světě (aniž by zatím tušil, jak velká jednou bude), připravil půdu pro pěkný zmatek. Řada článků se v různých jazykových mutacích fakticky lišila (nemuselo nutně jít o chyby – např. hodnoty mohly být počítány podle jiné metodiky). V roce 2012, kdy situace začala být fakticky neúnosná, přišli němečtí wikipedisté s ideou projektu **Wikidata.org**. Nešlo o nic menšího než databáze, ze kterých budou všechny jazykové mutace webové encyklopedie dynamicky získávat informace. Tím značně usnadnili rozvoj

projektu DBpedia.org („databázová encyklopedie“), který data z Wikipedie propojuje na další datasety. ←

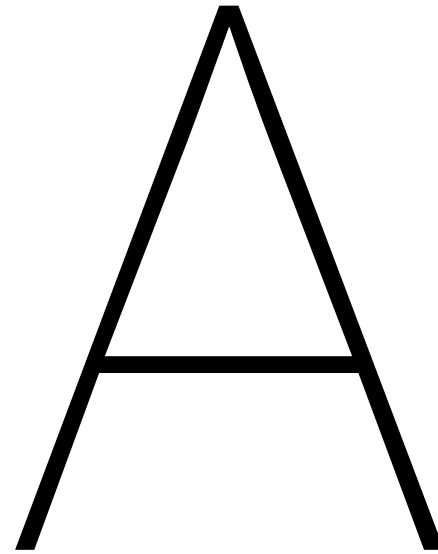
Další zdroje informací o propojených datech

Pokud vás ale zajímá, co se bude s daty dít dál, podívejte se na [Linkeddata.org](http://linkeddata.org). Doporučit můžeme též elektronickou knihu **Linked Open Data: The Essentials: A Quick Start Guide for Decision Makers** [48].

[46] Běžně se používá Linked Open Data (LOD).

[47] Anglicky se propojení více služeb označuje jako „mashup“, český překlad – „míchanice“ – se příliš nepoužívá.

[48] Kaltenböck, Martin Florian, Bauer. Linked Open Data: the Essentials: a Quick Start Guide for Decision Makers [online]. Wien: ed. mono/monochrom, 2012 [cit. 2014-06-08]. ISBN 978-390-2796-059. Dostupné z: <http://www.semantic-web.at/LOD-TheEssentials.pdf>.



**příloha:
definice
otevřených
dat**

Uveďme alespoň několik relevantních zdrojů, které považujeme za směrodatné:

Open Definition

Open Knowledge (OKFN) šíří definici „otevřených znalostí“, kterou považujeme za velmi obecnou (a tedy dobře přijatelnou) pro otevřená data. Otevřená ve stručnosti znamená, že data jsou k dispozici online a kdokoli je může využívat, měnit a sdílet bez jakýchkoli omezení. Podrobná otevřená definice (v češtině) visí na [webu](#). Všechna následující vymezení jsou konkretizací Open Definition.

Pětihvězdičkové schéma

[Sir Tim Berners-Lee](#), vynálezce webu, definoval pětiúrovňovou stupnici (jednotlivé stupně jsou rozlišeny počtem hvězdiček), přičemž za otevřená data považuje už ta s jednou hvězdičkou – tedy jakýkoli dokument, který je online. Tématu se věnujeme v kapitole Kvalita dat; nyní stačí říct, že opětovně použitelná data by měla (minimálně) být online, v běžných (a otevřených) formátech a strojově čitelná. Všechny pět úrovní detailně popisuje web [5stardata.info](#).

Open Data Index ^[49]

[Open Knowledge Foundation](#) spustila v roce 2012 přehled, jak jsou na tom státy s otevíráním důležitých datasetů. Projekt se od té doby profesionalizoval a byl doplněn o další země. Pro účely indexu bylo třeba naformulovat, které datasety jsou důležité a kde bude hranice otevřenosti.

Jako zásadní na základě zkušeností lídra tématu v OKF stanovili tyto datasety ^[50]:

- Jízdní řády
- Státní rozpočet
- Státní útraty

- Výsledky voleb
- Obchodní rejstřík
- Mapy států
- Základní statistiky
- Legislativa
- Poštovní směrovací čísla
- Emise škodlivin v ovzduší

U každého datasetu bylo hodnoceno 9 požadavků:

- Data existují (tj. dá se k nim dostat například přes zákon o svobodném přístupu k informacím)
- Data existují v digitální podobě (nejsou výhradně na papíru)
- Jsou veřejně dostupná
- Jsou dostupná zdarma
- Jsou online
- Jsou strojově čitelná
- Jsou dostupná vcelku
- Jsou pod otevřenou licenci
- Jsou aktuální

[49] Dříve Open Data Census.

[50] Výběr je samozřejmě arbitrární, do jisté míry ovlivněný projekty nad uvedenými daty, které OKFN provozuje. Je však i naší zkušeností, že právě o těchto 10 datasetů je i v Česku největší zájem (Seznam.cz dlouhodobě usiluje o jízdní řády, státní finance řeší projekt Budováním státu.cz, výsledky voleb zpracovává KohoVolit.eu, legislativu Zákonyprolidi.cz, emise škodlivin aplikace SmogAlarm.cz. Mapy, výstupy ČSÚ, PSČ a obchodní rejstřík používá nesčetně komerčních aplikací.

Zákon o svobodném přístupu k informacím a zákon o právu na informace o životním prostředí

Zákony 106/1999 Sb. a **123/1998 Sb.** byly první právní normy, které pracovaly s poskytováním informací ze strany státní správy a samosprávy občanům a které se dotýkaly též opakovaného použití informací.

Zákon o svobodném přístupu k informacím nařizuje povinným subjektům (což jsou veškeré instituce státní správy a samosprávy, dokonce organizace a firmy jimi zřízené a vlastněné) publikovat několik povinných informací a dále pak poskytovat informace na žádost občanů. Od roku 2012 žádosti zjednodušuje webová aplikace Infoprovsechny.cz.

Zákon v některých formulacích podporuje znovupoužití informací:

- Pokud je to možné s přihlédnutím k povaze podané žádosti a způsobu záznamu požadované informace, poskytnou povinné subjekty informaci v elektronické podobě.
- Povinné subjekty jsou povinny zveřejňovat informace uvedené v odstavci 1 a 2 též způsobem umožňujícím dálkový přístup.
- Pokud je informace zveřejněna v elektronické podobě, musí být zveřejněna i ve formátu, jehož specifikace je volně dostupná a použití uživatelem není omežováno.

Zákon se ale nezabývá konkrétní specifikací a poskytovatelé často ztěžují opětovné využití informace zhoršením jejich kvalit. Získávat otevřená data přes zákon 106/1999 Sb. je z dosavadní zkušenosti spíš nouzová možnost.

Autor platného zákona **Oldřich Kužilek** prosazuje v posledních letech novelu zákona, která se pokouší otevřená data zahrnout pod obecnější přístup k informacím. Podoba novely se ale stále mění, není proto relevantní uvádět k ní zde detaily.

Směrnice PSI

Směrnice Evropského parlamentu a Rady č. 2003/98/ES ze dne 17. listopadu 2003 o opakovaném použití informací veřejného sektoru zavedla otevřená data do evropské legislativy. Říká, že data produkovaná státní správou by měla být opětovně využitelná.

V roce 2014 parlament schválil novelizaci, která členskými státy nařizuje:

- Poskytovat veřejná data k jakýmkoli účelům znovuvyužitelným způsobem, zvláště přitom zmiňuje strojovou čitelnost.
- Držet se pravidla, že vládní data by měla být zdarma nebo za co nejmenší (a pevně stanovenou) cenu.
- Chápat jako data veřejného sektoru (tedy ta, na která se směrnice vztahuje) i data řady „paměťových“ institucí (jako jsou muzea a univerzitní knihovny).

Vlády zemí EU mají směrnici implementovat do konce roku 2015.

Metodika otevřených dat veřejné správy

Martin Nečaský, Jan Kučera a Dušan Chlapek v Metodice otevřených dat veřejné správy [51], psané pro Úřad vlády ČR, formulovali na základě principů z pera Sunlight Foundation [52] deset podmínek pro otevřená data v Česku. Na rozdíl od pětihvězdičkového schématu zde autoři předpokládají splnění všech šesti podmínek, aby data mohla být považována za otevřená:

[51] Chlapek, Dušan, Jan Kučera a Martin Nečaský. Metodika publikace otevřených dat veřejné správy ČR: verze 1.0. In: Boj s korupcí v České republice [online]. 2011 [cit. 2014-06-08]. Dostupné z: http://www.korupce.cz/assets/partnerstvi-pro-otevrene-vladnuti/otevrena-data/Metodika_Publ_OpenData_verze_1_0.pdf.

[52] <http://sunlightfoundation.com/policy/documents/ten-open-data-principles/popř.> <http://bit.ly/10principlesOGI>

- Úplná – data jsou zveřejněna v maximálním možném rozsahu. Rozsah může být definován právním předpisem, usnesením vlády, případně poskytovatelem dat.

- Snadno dostupná – data jsou dostupná na internetu a dohledatelná běžnými ICT nástroji a prostředky.

- Strojově čitelná – data jsou ve formátu, který je strukturován takovým způsobem, že z nich lze pomocí programové aplikace získat žádané (vybrané) údaje.

- Používající standardy s volně dostupnou specifikací (otevřené standardy) – data musí být ve formátu, který je volně (bezplatně) dostupný pro libovolné použití nebo do takového formátu převoditelný volně dostupnou aplikací.

- Zpřístupněná za jasně definovaných podmínek užití dat (licence) s minimem omezení – podmínky musí být jasně a zřetelně definovány a zveřejněny.

- Dostupná uživatelům při vynaložení minima možných nákladů na jejich získání – poskytovatelé jsou v souvislosti s poskytováním dat oprávněni žádat úhradu maximálně ve výši, která nesmí přesáhnout náklady spojené s jejich zpřístupněním uživateli; poskytovatel dat může jednorázově žádat i úhradu za mimořádně náročné pořízení dat, pokud si uživatel zpřístupnění těchto dat vyžádá.

Dále je vhodné (nikoli však nutné), aby otevřená data byla:

- Primární (původní) – data, která jsou zveřejněna původcem dat v podobě, v jaké byla původcem vytvořena. Za primární data se považují:

Referenční údaje ze základních registrů

Data z registrů a rejstříků veřejné správy

Agregovaná data (například výsledky voleb), pokud není možné zveřejnit data, z nichž byla provedena agregace

Agregovaná data (například statistiky nad jinými otevřenými daty), pokud je uveden způsob agregace a odkaz na zveřejněná primární data, z nichž byla agregace provedena

- Zveřejněná bez zbytečného odkladu – zveřejnění dat není zdrženo činnostmi, které nesouvisí s jejich přípravou; činnosti nezbytné pro publikaci dat jsou provedeny v čase, který umožní jejich zveřejnění bez nepřiměřeně dlouhé prodlevy od okamžiku vzniku dat

- Neomezující přístup – data dostupná způsobem, který nediskriminuje jednotlivce nebo skupinu osob

- Stále dostupná – data jsou dostupná online po dobu uvedenou jejich poskytovatelem

B

**příloha:
struktura dat**

Granty

Proč zveřejňovat

- Granty z veřejných prostředků mají nepochybně být dohledatelné a dostupné veřejnosti.
- V rámci přípravy zveřejňování dat je možné také racionalizovat oběh dokumentů a dat (cílem by měla být minimalizace přepisování dat žádosti do tabulky, z tabulky do systému, ze systému do publikovaného datasetu apod.).

Jaké datasety zveřejňovat

- Grantové okruhy a grantové programy [53]
- Výsledky grantové soutěže
- Vazba mezi granty a číslem zakázky/akce v ekonomickém systému – tím bude řešena vazba na rozpočet
- Vazba mezi granty a uzavřenými smlouvami o poskytování dotace

Jak zveřejňovat

- O fyzických osobách (žadatelích) publikovat následující údaje:
Jméno, příjmení, adresa (v žádném případě rodné číslo, číslo občanského průkazu nebo jiný jednoznačný – a zneužitelný – identifikátor)
- Do podmínek grantové soutěže dát souhlas žadatelů se zpracováním osobních údajů v rozsahu, který umožní jejich publikaci na webu [54]
- U historických dat, kde ještě není v podmínkách soutěže dostatečný souhlas se zpracováním osobních údajů, neposkytovat osobní údaje o fyzických osobách, jejichž projekty nebyly podpořeny

- Struktura výsledků grantové soutěže:

Žadatel
Adresa žadatele
Název projektu
Datum zahájení projektu
Datum ukončení projektu
Stručný popis – musí být součástí formuláře žádosti o grant
Závěrečná zpráva o realizaci – pouze u podpořených projektů
Evidenční číslo projektu – zvážit použití čísla jednacího ze spisové služby
Požadovaná dotace
Schválená dotace
Forma přihlášky a sběr dat

Formulář žádosti o grant obvykle města žadatelům poskytují elektronicky jako editovatelný soubor ve formátech MS Office. To však diskriminuje ty, kdo užívají konkurenční software a rovněž nutí administrátora žádosti kopírovat data do souhrnných tabulek. Webový formulář by mohl vést i k usnadnění procesu zpracování žádostí o granty.

- Přidělená výše finančních prostředků v datasetu Grantové programy nebude z důvodu toho, že zastupitelstvo se může rozhodnout, že přidělí prostředky jinak než v rámci předpokladáných výší.
- Zveřejňovat i samotné obsahy žádostí a ekonomické vyhodnocení a závěrečné zprávy projektů?

U nepodpořených projektů je riziko, že neúspěšnému žadateli na základě zveřejněné přihlášky do další grantové soutěže někdo nápad „ukradne“.

[B] Uvedené informace jsou převzaty z prezentací a zápisů workshopů Fóra pro otevřená data v roce 2014. Chlápek Dušan, Nečaský Martin a Kučera Jan. Prezentace pro workshopy se zástupci měst. Praha, 2014.

[53] Např. v Opavě vyhláší 9 grantových programů ve třech okruzích (sport, kultura, životní prostředí). V mnoha případech bude takové členění pro nižší počet grantových výzev/programů zbytečné.

[54] Zveřejnění jména a adresy v případě, že dotyčný obdrží grant z veřejného rozpočtu, je situace analogická úpravě zákona 106/1999 Sb., který u žádosti umožňuje poskytnout osobní údaje ve vymezeném rozsahu u příjemců veřejných prostředků. Protože se to ale nedá vztáhnout na neúspěšné žadatele, je třeba žádat explicitní souhlas.

Součástí podmínek grantové soutěže by mohl být souhlas se zveřejněním přihlášky a závěrečné zprávy a ekonomického vyhodnocení projektu.

U podpořených projektů by měla být zveřejněna závěrečná zpráva z projektu.

Asi postačí, když bude u každého projektu uveden jeho stručný popis – ten musí být součástí grantové přihlášky, aby bylo možné popisy do výsledné datové množiny jednoduše převzít.

● Vazba na rozpočty:

V rámci rozpočtu by měla být vyčleněna položka na granty, je možné, že to bude i členěno dle kapitol jako například sport, kultura. Každý grant by z hlediska rozpočtu měl mít vlastní identifikátor (číslo zakázky, příznak apod.). Například v Opavě jsou granty uvedeny pod číslem akce v analytickém účetnictví. Vazba na rozpočet bude nejlépe řešena samostatnou datovou množinou, ve které by byly výstupy z ekonomického systému o financování grantů.

Smlouvy [55]

● Doporučeno získat souhlas smluvních stran se zveřejněním smluv – jako součást ustanovení uzavřené smlouvy (pak nebude třeba data anonymizovat).

● Pokud budou publikovány smlouvy podepsané, bude patrně větší jistota, že zveřejněná smlouva bude odpovídat skutečnosti, ale je třeba sledovat vývoj v oblasti zákona o registru smluv, který tuto oblast může upravit.

● IČ subjektu není třeba přidávat, lze ho dohledat podle ID datové schránky.

● Struktura dat:

ID datové schránky – jeden orgán veřejné správy může mít více datových schránek.

Je potřeba, aby bylo možné zadat vícekrát subjekt „Partner“ – více smluvních stran.

Smlouvy, faktury a objednávky by měly být svázány jedním jednoznačným číslem.

Veřejné zakázky

● Doporučeno v max. míře využívat profily zadavatele a zajistit kvalitu tam publikovaných dat.

Akce, události

● Akce, které jsou z podpořených žádostí, by bylo možné na data o grantech propojit.

Do struktury metadat přidat garanta (kurátora) dat jako samostatný metadatový atribut.

[55]

Zveřejňování smluv bude možná povinné – v roce 2014 probíhá vzrušená debata o návrhu poslance Jana Farského podle slovenské předlohy. Návrh podporuje též iniciativa Rekonstrukce státu, která speciálně k tomuto zákonu zřídila web Nejkrásnější zákon <http://nejkrasnejsizakon.cz/>.

C

**příloha:
authority**

Vedle informačních zdrojů, které uvádíme dále, považujeme za užitečné vás nasměrovat na lidi, kteří mají v oblasti otevřených dat největší autoritu a o kterých je dobré vědět – ať už je budete sledovat na sociálních sítích, nebo se setrnete fyzicky.

Tuzemské

Fórum pro otevřená data

Fórum pro otevřená data je expertní program Fondu Otakara Motejla, Matematicko-fyzikální fakulty UK a Fakulty informatiky a statistiky VŠE v Praze. Zaměřuje se na vzdělávání vysokoškolských studentů, osvětu a medializaci otevřených dat a jejich využití v praxi. Poskytuje také obcím a institucím metodiku pro aplikaci otevřených dat a konzultace pro dlouhodobě udržitelná řešení v práci s daty.

Fórum tvoří:

Ing. Dušan Chlapek, Ph.D., je odborným asistentem na katedře informačních technologií Fakulty informatiky a statistiky VŠE v Praze. Ve své praxi se podílel na návrhu a implementaci řady informačních systémů pro organizace v soukromém sektoru i ve veřejné správě. Je jedním z autorů Koncepce katalogizace otevřených dat veřejné správy ČR a Metodiky publikace otevřených dat veřejné správy ČR.

Ing. Jan Kučera je studentem doktorského studia na katedře informačních technologií Fakulty informatiky a statistiky VŠE v Praze, kde působí také jako vědecký pracovník. Podílel se na tvorbě obsahu katalogu Cz.ckan.net a spolupracoval na výše uvedené koncepci a metodice.

Mgr. Martin Nečaský, Ph.D., je odborným asistentem na katedře softwarového inženýrství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze. Ve své práci se zaměřuje na technologie XML, propojitelná data (Linked Data) a otevřená propojitelná data (Linked Open Data). Taktéž se podílel na výše uvedené koncepci a metodice.

Mgr. Jiří Knitl je manažerem Fondu Otakara Motejla.

Mgr. Michal Kubáň, M.A., je advokátem otevřených dat ve Fondu Otakara Motejla. O otevřených datech publikuje na [Twitter.com/kubanster](https://twitter.com/kubanster).

FreeGeoCZ

FreeGeoCZ je otevřená mailová konference odborníků na geodata, ke které se můžete připojit na bit.ly/freegeoccz.

Jindřich Mynarz

Jindřich Mynarz (Mynarz.net) má v Česku nejširší zahraničí kontakty, sám je v mezinárodní komunitě respektován jako odborník na propojená data. V letech 2011 a 2012 organizoval akci BigClean (konference a workshopy zaměřené na zpracování dat). Aktivní je zejména na [Twitter.com/jindrichmynarz](https://twitter.com/jindrichmynarz).

Michal Berg

Michal Berg je v oblasti dat podnikatelem, aktivistou i politikem. Na Datablog.cz publikuje texty především o otevřených datech ve veřejné správě a jejich využití. Nejvíce obsahu sdílí na [Twitter.com/michalberg](https://twitter.com/michalberg).

Jiří Skuhrovec

Jiří Skuhrovec je programátor a ekonom, podílel se na projektech [Politickéfinance.cz](https://www.politickefinance.cz), [zindex.cz](https://www.zindex.cz) nebo [Vášmajetek.cz](https://www.vasmajetek.cz). Z veřejně angažovaných akademiků má největší zkušenosti se získáváním dat státní správy. Aktuální přehled jeho aktivit najdete na [LinkedIn](https://www.linkedin.com/in/jiri-skuhrovec).

Zahraniční

Tim Berners-Lee

Sir Tim Berners-Lee je původem britský fyzik, který stojí za **vznikem webu**. Dodnes se prostřednictvím organizací jako jsou W3C, The Web Foundation, Open Knowledge Foundation nebo The Open Data Institute, angažuje v rozvoji webu. Otevřená data jsou jeho konceptem. Tim příliš nepublikuje, ale vystupuje na řadě konferencí, vyplatí se sledovat třeba **TED Talks**.

Open Knowledge [56]

Open Knowledge (**Okfn.org**) je britská nezisková organizace, která celosvětově propaguje otevřené znalosti a otevřená data. Organizuje Open Knowledge Conference a Festival, stojí za katalogem CKAN nebo Open Data Index.

The Open Data Institute

Open Data Institute (**Theodi.org**) je britská neziskovka založená Timem Berners-Leem. Propaguje otevřená data zejména ve vládní a korporátní sféře. Její výzkum je podporován britskou vládou.

Andrew Stott

Andrew Stott byl ředitelem sekce pro transparentnost a digitální služby britské vlády. Stál za vytvořením národního katalogu otevřených dat Data.gov.uk ve Velké Británii. Dodnes je členem Rady vlády pro transparentnost; přednáší o implementaci otevřených dat. Aktivní je na **Twitter.com/DirDigEng**.

W3C

Konsorcium, založené a řízené Timem Berners-Leem (**w3c.org**), tvoří a rozvíjí technologické standardy pro web. V oblasti standardizace je práce W3C nezastupitelná.

Share-PSI 2.0 Project

Evropská síť pro sdílení informací ohledně implementace **PSI směrnice**.

Project Open Data na Githubu

Knihovna materiálů k otevřeným datům.

D

**příloha:
výkladový
slovník,
glosář**

Pokud by strohé popisy nestačily, velmi dobrý glosář nabízí například web [W3C](#).

API

Application Programming Interface; programátorské rozhraní pro přístup k datům; více na <http://cs.wikipedia.org/wiki/API>

Big Data

Obecně práce s velkými daty; od standardního sběru dat a jejich analýzy se Big Data liší především hardwarovými nároky (výrazně vyššími) a (obvykle) lepší predikcí na základě dosavadního průběhu; http://cs.wikipedia.org/wiki/Big_data.

CENIA (Czech Environmental Information Agency)

Státní instituce zřízená Ministerstvem životního prostředí ČR; provozuje český geoportál; <http://www1.cenia.cz/CKAN> – open source k tvorbě datového katalogu z produkce Open Knowledge; <http://ckan.org>.

CMS (Content Management System)

Software zajišťující správu obsahu, nejčastěji webu; termín je často zaměňován za redakční nebo publikační systém; <http://cs.wikipedia.org/wiki/CMS>.

Creative Commons

Sada licenčních ujednání, alternativa ke copyrightu. Jistá kombinace prvků CC licence je pokládána za vhodnou pro otevřená data; http://cs.wikipedia.org/wiki/Creative_Commons.

Databáze

Uspořádaná množina dat; laicky ji chápeme jako velkou a složitou tabulku; <http://cs.wikipedia.org/wiki/Datab%C3%A1ze>.

Datový model

Definuje strukturu dat v databázi/tabulce; laicky si ho vykládáme jako záhlaví tabulky.

Datový katalog

Strukturovaný web mající funkci rozcestníku na stránky jednotlivých datasetů.

ESRI Shapefile

Datový formát pro ukládání vektorových informací, typický pro geodata (např. hranice územních jednotek, obrysy parcel apod.); <http://cs.wikipedia.org/wiki/Shapefile>.

DDoS, DoS

Odmítnutí služby – reakce serveru na přehlcení požadavky ze strany skutečných nebo virtuálních uživatelů (tzv. dotazy); často zneužíváno k hackerským útokům; http://cs.wikipedia.org/wiki/Denial_of_service.

DMS (Documents Management System)

System zajišťující správu elektronických dokumentů, často používaný na úřadech v kombinaci se spisovou službou; http://cs.wikipedia.org/wiki/Document_Record_Management_System.

Fusion Tables

Viz Google Fusion Tables.

Geodata

Datasety mající reálnou kartografickou reprezentaci (je možné je zobrazit na mapě).

GML

Skriptovací programovací jazyk; <http://cs.wikipedia.org/wiki/Gml>.

Google Fusion Tables

Nástroj vyvinutý pro základní interpretace a vizualizace dat; https://support.google.com/fusiontables/answer/2571232?hl=en&ref_topic=1652595. Kurz, jak se s nástrojem naučit pracovat, je k dispozici na <https://datasense.withgoogle.com/course>.

Granularita

V případě dat znamená, jak velkou oblast zahrnuje jedna hodnota (například průměrná hodnota za minutu/za celý den) a jak je tato hodnota členěná; <http://en.wikipedia.org/wiki/Granularity>

HTML (HyperText Markup Language)

Jazyk, kterým jsou kódovány webové stránky; http://cs.wikipedia.org/wiki/HyperText_Markup_Language.

IoT (Internet of Things – internet věcí)

Technologický trend připojovat jakoukoli nevyčíslenou techniku (ledničky, auta, pračky...) k internetu. Obvykle se jako IoT označuje i masivní nasazení senzorů. Související pojmy: Big Data. Související nástroje: www.ifttt.com; http://en.wikipedia.org/wiki/Internet_of_Things.

IP adresa

Kód, kterým se jakékoli zařízení připojující se do sítě (notebook, tablet, telefon, tiskárna) jednoznačně identifikuje;
http://cs.wikipedia.org/wiki/IP_adresa.

JSON, GeoJSON

Otevřený datový formát nezávislý na počítačové platformě, hojně (ačkoli nejen tam) využívaný pro geodata; <http://cs.wikipedia.org/wiki/JSON>.

KML

Aplikace jazyka XML určená pro geodata;
<http://cs.wikipedia.org/wiki/KML>.

LOD, LOD Cloud

Linked Open Data, česky propojená otevřená data, jsou vysoce strukturovanou podobou datasetu s datovým modelem, který využívá URI, a je tak možné ho propojovat na jiné datasety.

Metadata

Strukturovaná data o datech; obsahují informace např. o obsahu datasetu, jeho tvůrci nebo licenci; <http://cs.wikipedia.org/wiki/Metadata>.

Microdata

Sémantická technologie v HTML5, dávající pouhým znakům (které mají smysl pro člověka, nikoli pro počítačový program) dávají význam; podobně RDFa; http://cs.wikipedia.org/wiki/HTML5_mikrodata.

OCR (Optical Character Recognition)

Technologie optického rozpoznávání znaků, software, který z fotografie nebo naskenovaného obrázku dokáže vytvořit strojově čitelná data (řetězec znaků); <http://cs.wikipedia.org/wiki/OCR>.

ODI (Open Data Institute)

Mezinárodní instituce se sídlem v Londýně mající za úkol osvětu a rozvoj otevřených dat ve světě; <http://cs.wikipedia.org/wiki/ODI>.

OGC (Open Geospatial Consortium)

Mezinárodní sdružení téměř 500 subjektů pracujících s geodaty;
<http://www.opengeospatial.org>.

OGP (Open Government Partnership)

Česky Partnerství pro otevřené vládnutí, mezinárodní iniciativa podněcující státy k otevřenějšímu vládnutí; ČR je zapojena od roku 2011; v rámci

tzv. Akčního plánu OGP se česká vláda poprvé zavázala k otevírání dat;
<http://opengovpartnership.org>.

OKFN (Open Knowledge)

Mezinárodní nezisková organizace působící v oblasti otevřených dat, stojící za řadou produktů, projektů a iniciativ (Open Data Index, CKAN, Where Does My Money Go?, Data-driven Journalism Handbook atd.);
<https://okfn.org>.

Ontologie

Explicitní popis problematiky; v této knize především dat; související pojmy: metadata, datový model;
[http://cs.wikipedia.org/wiki/Ontologie_\(informatika\)](http://cs.wikipedia.org/wiki/Ontologie_(informatika)).

OpenRefine

Freeware určený k čištění dat; <http://openrefine.org>.

Open Source

Software s otevřeným (tj. zkopírovatelným, opravitelným, doplnitelným) zdrojovým kódem; <http://bit.ly/otevrenysoft>.

Otevřený formát

Formát, který je čitelný pro open source software.

Parser

Algoritmus převádějící řetězec znaků (text) do strukturované podoby (tabulky); nástroj užíván zvláště pro scrapování;
<http://bit.ly/syntaktanaliza>.

Prostorová data

Viz geodata.

PSI (Public Sector Information)

Informace veřejného sektoru, někdy označovaná jako „Government Data“; související směrnice EU se někdy označuje jako PSI – směrnice.

RDF, RDFa

Technologický nástroj pro přenos strukturovaných informací; jeden ze základních (a konsorciem W3C doporučených) prvků sémantického webu; <http://en.wikipedia.org/wiki/RDFa>.

Re-use

Obecný pojem pro znovuvyužití (secondhandy, recyklace apod.); v souvislosti s daty se mluví o takové podobě datasetů, aby byly co nejsnadněji znovuvyužitelné. Re-use je pro otevřená data klíčovým termínem.

RÚIAN

Registr územní identifikace a nemovitostí; bit.ly/registrRUIAN.

Scraper

Počítačový program extrahující data z (nejčastěji) webových stránek a ukládající je ve strukturované podobě; http://en.wikipedia.org/wiki/Data_scraping.

Semantic web

Soubor standardů publikovaných konsorciem W3C, který má vést ke „strojově srozumitelnému“ webu a propojování informací. Klíčovou technologií je RDF. http://en.wikipedia.org/wiki/Semantic_Web.

Směrnice EU

Směrnice o opakovaném využití informací ve veřejném sektoru, plné znění na bit.ly/smernice, srozumitelně o tom na bit.ly/DATABLOG a <http://www.datablog.cz/clanky/psi>.

Stošestka

Viz zákon 106/1999 Sb.

Tim Berners-Lee

Britský informatik, vynálezce webu, ředitel konsorcia W3C; http://cs.wikipedia.org/wiki/Tim_Berners-Lee.

UIS-ADR

Územně identifikační registr adres; bit.ly/formsmpsv.

ÚOOÚ (Úřad pro ochranu osobních údajů)

Český správní úřad, který vykládá zákon o ochraně osobních údajů; jeho rozhodnutí bývají zmiňována jako překážky pro otevřená data; <http://www.uouu.cz>.

URI (Uniform Resource Identifier)

Obecnější termín, zahrnující (a často také slučovaný s) URL.

URL (Uniform Resource Locator)

Řetězec znaků jednoznačně identifikující dokument na webu; laicky řečeno: webová adresa; <http://en.wikipedia.org/wiki/Url>.

W3C (World Wide Web Consortium)

Mezinárodní konsorcium vyvíjející standardy pro web (HTML, RDF ad.); ředitelem je Tim Berners-Lee; <http://www.w3.org>.

WCS (Web Coverage Service)

Standard pro reprezentaci geodat; http://en.wikipedia.org/wiki/Web_Coverage_Service

WFS (Web Feature Service)

Standard zajišťující vzájemnou komunikaci mezi geodatovými aplikacemi na webu; http://en.wikipedia.org/wiki/Web_Feature_Service.

WMS (Web Map Service)

Standardní protokol pro pulikaci geodat na základě GIS; http://en.wikipedia.org/wiki/Web_Map_Service.

XML (Extensible Markup Language)

Značkový jazyk, jakási „nadstavba/rozšíření HTML“, oproti kterému se odlišuje strojovou čitelností; http://cs.wikipedia.org/wiki/Extensible_Markup_Language.

Zákon 106/1999 Sb.

Zákon o svobodném přístupu k informacím <http://www.zakonyprolidi.cz/cs/1999-106>.

Zákon 123/1998 Sb.

Zákon o právu na informace o životním prostředí <http://www.zakonyprolidi.cz/cs/1998-123>.

Zákon 365/2000 Sb.

Zákon o informačních systémech veřejné správy <http://www.zakonyprolidi.cz/cs/2000-365>.

E

**příloha:
seznam
zdrojů**

Sociální média, weby a e-mailové konference

Twitter: [#opendatacz](#)
[Česká OKFN e-mailová konference](#)
[Otevenadata.cz](#)
[Opendata.cz](#)
[Mvcr.cz/clanek/otevrena-data-aspx](#)

Příručky, kurzy

United Nations. Guidelines on Open Government Data for Citizen Engagement [online]. New York: United Nations, 2013 [cit. 2014-05-22]. Dostupné z: <http://workspace.unpan.org/sites/Internet/Documents/Guidelines%20on%20OGDCE%20May17%202013.pdf>.

World Bank. Open Government Data Toolkit [online]. 2012 [cit. 2014-05-22]. Dostupné z: <http://data.worldbank.org/open-government-data-toolkit>.

Open Knowledge. The Open Data Handbook [online]. 2011 [cit. 2014-05-22]. Dostupné z: <http://opendatahandbook.org>.

Grey, Jonathan, Bounegru a Lucy Chambers. Data Journalism Handbook [online]. 2012 [cit. 2014-05-22]. Dostupné z: <http://datajournalismhandbook.org/1.0/en>.

D7.1.3 – Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC [online]. [cit. 2014-05-24]. Dostupné z: <https://joinup.ec.europa.eu/sites/default/files/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf>.

Kaltenböck, By Florian Bauer and Martin. Linked Open Data: The Essentials: a Quick Start Guide for Decision Makers [online]. Wien: ed. mono/monochrom, 2012 [cit. 2014-06-08]. ISBN 978-390-2796-059. Dostupné z: <http://www.semantic-web.at/LOD-TheEssentials.pdf>.

Tactical Tech. Drawing by Numbers [online]. [cit. 2014-06-08]. Dostupné z: <https://drawingbynumbers.org>.

Hellerstein, Joe a Amit Deutsch. Making Sense of Data [online]. 2014 [cit. 2014-06-08]. Dostupné z: <https://datasense.withgoogle.com/course>.

Government Digital Service. Government Service Design Manual: Build Services So Good that People Prefer to Use Them [online]. 2012 [cit. 2014-06-08]. Dostupné z: <https://www.gov.uk/service-manual>.

Government Digital Service. Design Principles [online]. 2012. vyd. 2012 [cit. 2014-06-08]. Dostupné z: <https://www.gov.uk/design-principles>.

Michael, Hausenblas. 5 star Open Data [online]. 2012 [cit. 2014-06-08]. Dostupné z: <http://5stardata.info>.

Lane, Kin. API Evangelist [online]. [cit. 2014-06-08]. Dostupné z: <http://apievangelist.com>.

Vládní dokumenty

Chlapek, Dušan, Jan Kučera a Martin Nečaský. Metodika publikace otevřených dat veřejné správy ČR: verze 1.0. In: Boj s korupcí v České republice [online]. 2011 [cit. 2014-06-08]. Dostupné z: http://www.korupce.cz/assets/partnerstvi-pro-otevrene-vladnuti/otevrena-data/Metodika_Publ_OpenData_verze_1_0.pdf.

Chlapek, Dušan, Jan Kučera a Martin Nečaský. Koncepce katalogizace otevřených dat VS ČR: zkrácená verze. In: Boj s korupcí v České republice [online]. 2011 [cit. 2014-06-08]. Dostupné z: <http://www.korupce.cz/assets/partnerstvi-pro-otevrene-vladnuti/otevrena-data/Koncepce-katalogizace--otevrenych-dat-VS-CR---zkracena-verze.pdf>.

Licence

Myška, Matěj, Libor Kyncl, Radim Polčák a Jaromír Šavelka. Veřejné licence v České republice. Brno: Masarykova univerzita, 2012. ISBN 978-80-263-0344-2. Dostupné z: <http://is.muni.cz/www/102870/Prirucka.pdf>.

<http://ict-law.blogspot.cz>.

Reporty

Chlapek, Dušan, Jan Kučera, Martin Nečaský a Michal Kubáň. Open Data and PSI in the Czech Republic. [online]. 2014 [cit. 2014-05-24]. Dostupné z: <http://www.epsiplatform.eu/content/open-data-and-psi-czech-republic>.

Citovaná literatura

Gleick, James. Informace: historie, teorie, záplava. 1. vyd. v českém jazyce. Praha: Dokořán, 2013, 396 s. Zip (Argo: Dokořán). ISBN 978-80-7363-415-5.

Thaler, Richard H a Cass R Sunstein. Nudge (Šťouch): Jak postrčit lidi k lepšímu rozhodování o zdraví, majetku a štěstí. Vyd. 1. Zlín: Kniha Zlín, 2010, 309 s. Tema (Kniha Zlín). ISBN 978-80-87162-66-8.

Johns, Adrian. Pirátství: boje o duševní vlastnictví od Gutenberga po Gatace. 1. vyd. Překlad Lucie Chlumská, Ondřej Hanus. Brno: Host, 2013, 633 s. ISBN 978-807-2947-119.

Braybrooke, Kaitlyn, Jussi Nissilä a Timo Vuorikivi. The open book [online]. London: Finnish Institute, 2013 [cit. 2014-06-08]. ISBN 978-0-9570776-2-1. Dostupné z: http://issuu.com/finnish-institute/docs/theopenbook_issuu_final.

Kašpárek, Michal. Slovníček pro desátá léta: (díl 1: od 3D tisku po GTD). 067.cz [online]. 2013, roč. 1, č. 1 [cit. 2014-05-24]. Dostupné z: <https://067.cz/archiv/1/porozumej-svemu-geekovi-slovnicek-pro-desata-leta-dil-1-2.html>.

Kundra, Vivek. Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect. [online]. 2012 [cit. 2014-05-24]. Dostupné z: http://shorensteincenter.org/wp-content/uploads/2012/03/d70_kundra.pdf.

Mráček, Jakub. Kdo vydělává na monopolu na státní data?. Lupa.cz [online]. 2013 [cit. 2014-05-24]. Dostupné z: <http://www.lupa.cz/clanky/jakub-mracek-kdo-vydelava-na-monopolu-na-statni-data>.

Fond Otakara Motejla. Otevřená data v České republice: doporučení. 2014, 4 s.

Píповá, Tereza. Výdaje na veřejnou správu měst. (Diplomová práce) Praha, VŠE, 2013.

Tajtl, Martin. Otevřená data o kontrolách ČOI. In: [online]. 2014 [cit. 2014-05-24]. Dostupné z: http://www.issc.cz/archiv/2014/download/prezentace/coi_tajtl.pdf.

Informační dopravní systém versus otevřená data. Smart Cities. 2014, 1., č. 1, 28–33.

Zajíček, Petr. Vliv příjmové politiky města na jeho výdaje. (Diplomová práce) Praha, VŠE, 2013.

ISBN 978-80-87725-15-3

editor
jakub mráček

Vydal Fond Otakara Motejla
v roce 2014.

autoři textů
jan boček
jáchym čepický
jakub mráček

Jak otevírat data, jehož autorem je
Mráček Jakub, podléhá licenci
Creative Commons.
Uvedte autora a zachovejte licenci
4.0 Mezinárodní.

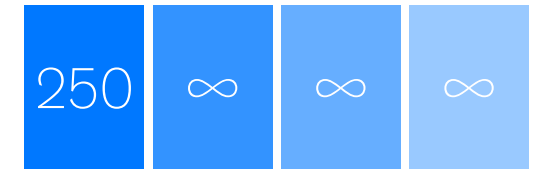
grafická úprava
ex lovers

korektury



čtení revize

formáty



tisk pdf epub mobi