Český překlad http://opendatahandbook.org/

Překládáme Open Data Handbook, protože chceme, aby se otevřená data stala standardem v české veřejné správě i mezi neziskovkami. Záměrně jako první překládáme technickou část příručky (http://opendatahandbook.org/en/how-to-open-up-data/index.html), protože právě v této oblasti jsme nejčastěji žádáni o konzultaci. Výhledově ale dojde na kompletní překlad.

Následující hrubý překlad vytvořil stážista naší nadace, jeho úpravy proběhnou za přispění dobrovolníků a dobrovolnic. Finální verze překladu bude volně přístupná. Budeme potěšeni, pokud formou komentáře (Ctrl+Alt+M) přispějete k faktické správnosti a pořádné češtině!

Díky!
Jakub Mráček, @jakubmracek, jakub.mracek@osf.cz

Jak zpřístupnit data

Tato kapitola je jádrem příručky, jelikož obsahuje konkrétní rady jak zpřístupnit data. Čtenář se zde dozví vše od základních informací až po velice delikátní problémy, které by se mohli v průběhu zmíněného procesu objevit.

Hned na začátek je důležité zmínit tři základní pravidla, která by měla být dodržována při zpřístupňování dat:

- **Jednoduchost.** Pokud se chystáte začít zpřístupňovat informace, začněte s malými objemy dat, které bude možné jednoduše a rychle zveřejnit. Není nutné, aby byla veškerá data zveřejněna okamžitě po spuštění celého projektu (je však pravdou, že čím více dat se vám zpočátku podaří zveřejnit, tím lépe).

  Celý proces zpřístupňování dat je jedna velká inovace, proto je důležité, aby se projekt rychle rozvíjel, čímž dojde k postupnému nabalování. Každá inovace sebou nese jak úspěchy, tak i nezdary – je možné, že některé zveřejněné informace nebudou použitelné, nicméně to by vás nemělo odradit.

- **Zveřejňujte data brzy a často.** Poskytujte současným i potenciálním uživatelům aktuální informace, což výrazně zvýší relevanci vámi poskytnutých dat (a je jedno, zdali jsou informace určené občanům, podnikatelům nebo vývojářům).

  Je nutné si uvědomit, že většina zpřístupněných informací nebude použita koncovými uživateli, ale spíše zprostředkovateli informací. Jedná se o lidi, kteří zpracují informace do srozumitelné podoby a následně je poskytnou koncovým uživatelům. Jako příklad může sloužit GPS – sami chcete požívat raději mapu než složitý systém údajů o vaší poloze, směru apod. Z tohoto důvodu zpřístupňujte data pro zprostředkovatele informací a nikoliv koncové uživatele, neboť ti dostanou již zpracovaná data.

- **Připravte se na obavy, omyly a fámy.** Pravidlo je velice důležité zvlášť ve státním sektoru.

Pokud budete zpřístupňovat data, narazíte na velké množství otázek a obav spojených s otevřenými daty. Z tohoto důvodu je důležité (a) vybrat ty nejnaléhavější (b) velice rychle je vyvrátit.

V rámci samotného procesu zpřístupňování dat můžeme rozlišit čtyři oddělené fáze vyjmenované níže. Jednotlivé fáze nejsou řazeny, tak jak by měly po sobě následovat – mnoho z nich může být prováděno současně.

1. **Vyberte si datový balíček(y) k zpřístupnění.** Je možné, že se k této fázi budete muset vrátit, pokud narazíte na problémy v pokročilejších fázích zpřístupňování
2. **Použijte otevřenou licenci.**
    a. Zjistěte, jaký typ duševního vlastnictví obsahují vámi zvolená data
    b. Použijte licenci, která je vhodná pro všechna data v balíčku a zároveň splňuje definici otevřenosti dat, tak jak bylo popsáno v kapitole „Co jsou otevřená data“
    c. Upozornění: pokud nejste schopni najít vhodnou licenci, vraťte se ke kroku 1 a zkuste použít jiný datový balíček
3. **Zpřístupněte data.** Zpřístupněte data v celku a v použitelném formátu, přičemž jako doplněk můžete data zveřejnit pomocí API.
4. **Udělejte data snadno dohledatelná.** Vytvořte na webu například centrální katalog se seznamem vámi zpřístupněných dat.

Použití otevřené licence (právní otevřenost)

V České republice je duševní vlastnictví chráněno autorským zákonem. Aby bylo možné Vámi zveřejněná data legálně použít, je nezbytné ke každé sadě přidělit licenci.

Jaké typy licencí můžete použít? Doporučujeme použít licence pro otevřená data, která jsou v souladu s definicí otevřených znalostí a zároveň jsou označeny jako vhodné pro vámi zpřístupňovaná data. Seznam licencí společně s vysvětlením jejich požití je možné nalézt na této internetové stránce:

- http://opendefinition.org/licenses/

Krátký jednostránkový manuál k použití licence otevřených dat je k dispozici na stránce Open Data Commons :

- http://opendatacommons.org/guide/

Zpřístupňování dat (technická otevřenost)

Otevřená data musí být přístupná jak z právního, tak z technického hlediska, dosažitelná v celku a ve formátu vhodném pro další strojové čtení. Proto by měly mít tyto vlastnosti:

Přístupné:

Informace by měli být přístupné za rozumnou cenu - měly by být volně přístupné na internetu formou „stažení zdarma“. Není tedy vhodné používat předplacená datová úložiště. Pokud se zpřístupňuji „otevřená“ data, měla by být dostupná zdarma.

Kompletní:

> Informace by měly být přístupné jako kompletní set. Pokud tedy musíte nějaká data zveřejňovat ze zákona, měla by být k dispozici ke stažení jako celek (API nestačí).

Otevřené pro další automatické zpracování:

> Je důležité, aby veškeré informace byly poskytovány v pro programy snadno čitelném formátu. Takový PDF soubor je vhodný pro kvalitní tisk nebo lidské oko, ale pro programy je těžko čitelný, přitom zprostředkovatelé informací často používají své vlastní programy pro vyhledávání informací.

Při zpřístupňování dat je dobré se držet těchto zásad:

- Jednoduchosti
- Rychlosti zpřístupnění nových dat
- Věcnosti

Především je lepší poskytnout surová data ihned, než perfektně zpracovaná po šesti měsících.

Nejlepším způsobem, jak zpřístupnit data, je internet. Úřady nejčastěji zveřejňují svá data na svých internetových stránkách nebo v centrálním katalogu, který odkáže tazatele na jím hledané informace. I přesto však existují alternativy.

Tyto alternativy jsou vhodné, pokud máte vy nebo uživatel nekvalitní připojení nebo data jsou příliš velká. V následující kapitole budou popsány možnosti, jak zveřejňovat velké objemy dat a přitom neplatit neúměrně velké náklady.

Online metody

Pomocí API (Application Programming Interface)

API se staly v oblasti zveřejňování dat v poslední době velice populárními. Umožňují programátorům vybrat žádanou část dat, přičemž zároveň nemusejí zveřejňovat celý balík informací. API bývá typicky připojena k databázi, která je aktualizována v reálném čase, což umožňuje API získávat okamžitě nejnovější informace.

Veškerá otevřená data by měla být zveřejňována v celku jako surové informace. S ohledem na to, bývají s API spojeny některé problémy, které je nutné brát v potaz:

1. Cena. Na vývoj a údržbu API je nutno vynakládat mnohem více peněz než na samotné zpřístupňování dat.
2. Očekávání. Aby byla API využívaná, musí být spolehlivá. Pokud se vyskytne v programu chyba, očekávají uživatelé, že ji úřad opraví, což je spojeno s dalšími náklady.

Zveřejnění surových dat umožňuje:

1. Stálý přístup k informacím i v okamžiku, kdy na vývoj API dojdou peníze, nebo úřad zruší celý projekt apod.
2. Snadný přístup ke všem datům, ze kterých si pak zpracovatel vybere ty, které potřebuje. Díky tomu, že úřad nijak data pomocí API nezpracovává, se výrazně sníží náklady na

zpřístupňování informaci a potažmo pravděpodobnost selhání ze strany úřadu.

3. Kdokoliv může vytvořit vlastní API dle libosti a nemusí se bát, že přestane mít data k dispozici, čímž veškeré náklady na vývoj vyjdou vniveč.

Zveřejňování dat vcelku umožní uživatelům použít je k účelům, pro které nebyly původně zveřejněny. Díky tomu mohou být konvertovány do jiných formátů (např. čísla převedena do grafů), spojeny s daty z jiných zdrojů, či ukládány na rozmanitých místech. Z těchto důvodů by měly být v případě použití API data přístupná dvěma způsoby – nejnovější pomocí API a zbylá v celku v jednom balíku.

Například Eurostat zpřístupňuje veškerá data v celku ve formátu TSV (Tab-separated values) nabízející 4000 datových souborů, který je aktualizovaný dvakrát denně. Vše je zveřejňované zároveň s dokumentací o náročnosti stahování a důleřitosti dat v souborech.

Dalším příkladem může být datový katalog patřící District of Columbia, který umožňuje stáhnout data jak v CSV či XLS formát, tak i jako „live feed".

Formáty zveřejňovaných souborů - přehled formátů pro použití

JSON

JSON je formát, který je snadno čitelný pro jakýkoliv programovací jazyk a potažmo tedy snadněji zpracovatelný pro počítače než například XML.

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

XML je široce používaný formát pro výměnu dat, jelikož používá jasně stanovenou strukturu dat, čímž umožňuje vývojářům přidávat informace, aniž by ohrozily čitelnost dokumentu.

RDF

RDF je formát doporučený W3C. Umožňuje zobrazovat data ve formě, která je vhodná 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 propojených otevřených dat ve Španělsku a Velké Británii. Vynálezce webu Tim Berners-Lee navrhl nedávno pěti-hvězdičkové schéma, které obsahuje RDF jako jeden z cílů projektů otevřených dat.

Tabulkový procesor

Mnoho úřadů má své informace uchované v tabulkových procesorech (například 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.

Comma Separated Files

CVS 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 – je v zásadě nemožné přijít na to co jednotlivá buňka znamená. Z tohoto důvodu je důležité k CVS souborům vždy přikládat příslušnou dokumentaci.

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

Textový dokument

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

Obecně je doporučeno použít pro zpřístupnění strukturovaných informací jiný formát než textový dokument.

Čistý text

Čisté textové dokumenty (.txt) jsou velice dobře strojově čitelné. Problém však je, že obecně neobsahují informace o struktuře zapsané přímo do dokumentu, což nutí programátory vytvořit tzv. praser, který dokáže data interpretovat.

Určité problémy může také způsobovat výměna „.txt“ formátu mezi operačními systémy – např. MS Windows, Mac OS X a jiné Unix systémy mají vlastní způsoby jak říci počítači, že dosáhli konce dokumentu.

Naskenované obrázky

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

Individuálně vytvořené formáty

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

HTML

V dnešní době jsou informace běžně zobrazovány v HTML. Tento formát je obzvláště užitečný pokud se informace často nemění a mají přiměřený objem. Pochopitelně je lepší zpřístupňovat data ve formátu jednodušším pro stahování apod., ale vzhledem k nákladům potřebným pro vytvoření HTML, se jedná o dobrou cestu, jak začít s celým procesem zveřejňování dat.

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 (http://developer.yahoo.com/yql/), přičemž jeho

účinnost se zvyšuje, pokud jsou data řádně označená.

Otevřené formáty

I když jsou informace zveřejňovány elektronicky na internetu ve strojově čitelném formátu, mohou být k ničemu díky problémům spojeným přímo s typem formátu.

Formáty, ve kterých jsou informace publikovány – jinými slovy způsob, jakým jsou digitálně uchovány – mohou být buď uzavřené, nebo otevřené. 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.

Formát je „uzavřený“, pokud se nemá zveřejněnou dokumentace, nebo pokud je z důvodu ochrany duševního vlastnictví jeho použití omezeno, i když k němu byla zveřejněná dokumentace. Zveřejnění informací v uzavřeném formátu může způsobit výrazné překážky pro jejich další využití, neboť nutí uživatele si zakoupit 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 vlastnících licencí formátů. Pokud přesáhnou náklady na přístup k formátu únosnou mez, může v nejhorším případě dojít k nepoužití zveřejněných informací.

V případě, že chce veřejný sektor svoje data zpřístupnit, měl by tak učinit **v otevřených a strojově čitelných formátech.**

Příklad: Informace o silniční dopravy ve Velké Británii

Programátor Andrew Nicolson se účastnil protestů proti stavbě nového obchvatu okolo Westbury. V rámci přípravy kampaně se pokoušel získat původní data o intenzitě silničního provozu, které užívali lidé propagující stavbu. Na základě práva na svobodný přístup k informacím se mu podařilo získat jisté relevantní informace, nicméně místní samospráva mu poskytla informace ve specifickém formátu, který přečetl pouze program na výpočet hustoty provozu od firmy Saturn. Vzhledem k tomu, že neexistovala žádná verze programu umožňující pouhé prohlížení shromážděných dat („read only version“), byla Andrewa skupina aktivistů nucena zakoupit si program za 600€ (a to pouze díky tomu, že software koupili jako výukový). Normálně se dal program pořídit za více jak 15 000€, což je mimo absolutně mimo možnosti normálního člověka.

Ačkoliv právo na svobodný přístup k informacím nezaručuje poskytnutí informací v otevřeném formátu, čím dál více států přijímá povinnost poskytovat data v otevřených formátech. Zásadním zlomem bylo rozhodnutí Obamovy administrativy v roce 2009, které řeklo, že:

V nejvyšší možné míře za současně platných předpisů by měly úřady publikovat informace v otevřených formátech tak, aby je bylo možné získat, stáhnout, označit a vyhledat pomocí běžně používaných webových aplikací. Otevřeným formátem se myslí takový, který je „nezávislý“, strojově čitelný a bez jakýchkoliv dalších restrikcí, které by omezily použitelnost v něm obsažených informací.

Webové služby

Webové služby jsou vhodné pro data, která se často mění a zároveň jsou omezená na objemu. Mezi nejpoužívanější způsoby, jak vytvořit webovou službu, se řadí SOAP a REST.

Databáze

Databáze umožňují (stejně jako webové služby) přístup k informacím v reálném čase. Mají však výhodu v tom, že si uživatelé mohou najít pouze ty části informací, které je zajímají.

Je nutné si však uvědomit, že umožňovat online přístup k databázi může být i nebezpečné (mohou obsahovat citlivé informace, o kterých nevíme). Z tohoto důvodu je dobré poskytovat online přístup pouze k databázím, které obsahují dobře zdokumentované informace. Další řešením může být zkombinování webové služby a databáze – zveřejníme pouze informace, které jsou nezávadné a přitom jsou z databáze (a nejedná se o nijak nákladné řešení).