



ZÁPADOČESKÁ
UNIVERZITA
V PLZNI

FAKULTA APLIKOVANÝCH VĚD

LINKED DATA, OPEN DATA a BIG DATA

Bc. Zuzana LACINÁ

Semestrální práce z předmětu Prostorové databáze – obor Geomatika

Zadavatel: **Ing. Karel JANEČKA, Ph.D.**

Katedra: **Katedra matematiky, oddělení geomatiky**

Akademický rok: **2013/2014**

OBSAH

1 LINKED DATA.....	3
1.1 Co jsou linked data?	3
1.2 Základní pravidla pro linked data.....	3
1.3 Definice URI, HTTP, SPARQL a RDF	4
1.4 Vstupní data pro publikování linked dat.....	7
1.5 Web of data - linked data	9
1.6 Topologie datových sad.....	11
1.7 Metadata	12
1.8 Linked data aplikace.....	12
1.9 Ochrana osobních údajů	16
1.10 Příklady konferencí a vybrané workshopy na téma linked data	16
1.11 Zajímavé video popisující linked data.....	16
2 OPEN DATA	17
2.1 Co jsou open data?	17
2.2 Otevřené formáty souborů.....	18
2.3 Publikování open dat.....	21
2.4 Příklady užití vládních open dat v USA.....	22
2.5 Veřejná správa jako zdroj dat	23
2.6 Proces publikace otevřených dat	25
2.7 Open data v ČR	26
2.8 Datové katalogy.....	26
2.9 Film Open data	29
3 LINKED DATA versus OPEN DATA	30
4 BIG DATA	31
4.1 Co jsou Big data?	31
4.2 Základní charakteristiky	33
4.3 Způsoby zpracování velkého objemu dat	35
4.4 Příklad Big Dat.....	37
4.5 Big data a GIS	37
4.6 Uživatelé Hadoop	39
4.7 Konference Big data	39
5. REFERENCE	40

1 LINKED DATA

1.1 Co jsou linked data?

Linked data popisují způsob zveřejňování strukturovaných dat, která mohou být vzájemně propojena a stávají se užitečnými. Jsou postavena na standardních webových technologiích, jako je protokol HTTP, formát RDF a identifikátor URI. Rozšiřuje se sdílení informací způsobem, který lze číst automaticky pomocí počítačů, a tím se umožňuje zobrazení dat z různých zdrojů [1].

Pojem linked data uvedl v roce 2006 **Tim Berners-Lee**, zakladatel World Wide Webu a později také myšlenek sémantického webu, jehož jsou linked data součástí. Tim-Berners Lee definuje sémantický web jako přístup, umožňující vyjádření informací ve strojově čitelné podobě. Jeho hlavní myšlenkou je provázání dat s využitím odkazů, což umožní lidem i strojům procházet související informace neboli linked data. Termín linked data nebývá překládán a používá se v původním anglickém znění, ale v česky mluvícím prostředí se řídce užívá ekvivalentu "propojená data" [2].

Linked data se odkazují na soubor osvědčených postupů pro poskytování a připojení strukturovaných dat na webu. Tyto osvědčené postupy byly přijaty v průběhu posledních několika let, rostoucím počtem poskytovatelů dat, což vede k vytvoření globálního datového prostoru obsahujícího miliardy webových dat.

Dodržování standardů a osvědčených postupů, které jsou základem těchto zásad, umožňuje efektivitu, interoperabilitu dat a opětovné použití dat na webu [1].

Cílem linked dat je volné zpřístupňování dat ve formě, která je jednoduše automatizovaně zpracovatelná [3].

Tim Berners-Lee formálně definoval **4 základní pravidla**, pro zveřejňování údajů na webu tak, aby se veškerá publikovaná data stala součástí jediného globálního datového prostoru:

1.2 Základní pravidla pro linked data

1. Používat URI jako globální identifikátor objektů.

Všeobecně není snadné vytvářet identifikátory, které budou jedinečné napříč celým webem. Nejvhodnějším nástrojem pro tyto účely jsou právě URI, tedy způsob dobře osvědčený už ze současného webu dokumentů.

2. Používat HTTP URI, aby byly objekty vyhledatelné na webu.

Druhé pravidlo vychází z prvního a navíc využívá další výhodu URI, a sice jeho kombinaci s HTTP protokolem. Pokud už je objekt identifikován svým unikátním URI, není problém učinit toto URI skutečně vyhledatelné na webu.

3. Pokud se někdo podívá na URI, je třeba poskytnout užitečné informace s využitím standardů.

Další pravidlo hovoří o poskytování užitečných informací za pomoci standardizovaných nástrojů. Mezi tyto nástroje patří v první řadě formát RDF a dotazovací jazyk SPARQL.

4. Používat odkazy na další URI, což umožní vyhledávat související objekty.

Poslední princip vyžaduje uplatnění externích RDF odkazů na další objekty, což umožní skutečné zapojení informace do světa linked dat. Obohacením dat o co největší počet takových odkazů splňujících sémantický význam se výrazně zvýší hodnota a smysl celé informace.

1.3 Definice URI, HTTP, SPARQL a RDF

Web nám umožňuje spojit související dokumenty. Stejně tak nám umožňuje propojit související data. Klíčové technologie, které podporují linked data, jsou **URI, HTTP, RDF a SPARQL** [1].

1. Uniform Resource Identifier

URI jsou globální unikátní identifikátory, které umožňují věci jednoznačně identifikovat v rámci jmenného prostoru celého webu. Tím je umožněno jednoduché odkazování na pojmenovanou věc.

Pro příklad: Absolutní URI: `<http://example.com/uri>` , Relativní URI: `<uri>`.

Poskytovatelé dat si mohou vybrat mezi dvěma vzory identifikátorů, pro použití k identifikaci entit, je to identifikátor **303 URI** a **Hash URI** [10].

303 URI

Jedním z těchto způsobů je přesměrování pomocí HTTP odpovědi 303. Jakmile je dereferencováno URI objektu, server odpoví s HTTP kódem 303 a vrátí URI dokumentu popisující příslušný objekt. Výraznou nevýhodou tohoto způsobu je, že k získání dokumentu jsou třeba dva HTTP požadavky. Na druhou stranu je však možné, pro každý objekt vytvořit příslušný dokument a nastavit jednotlivá přesměrování v konfiguraci serveru [4].

Hash URI

Pro identifikaci objektu se v rámci URI použije znak # (tzn. hash) takovým způsobem, že část před hashem odpovídá URI dokumentu o objektu a vytváří jmenný prostor pro podobné objekty. Část za hashem, fragment, obsahuje jedinečný identifikátor objektu v rámci tohoto jmenného prostoru.

Před odesláním na server je fragment vždy odstraněn a server tak rovnou obdrží požadavek na URI dokumentu. Použitím tohoto způsobu stačí jen jeden HTTP požadavek na server. Nelze však na serveru rozlišit, na který konkrétní objekt byl vznesen požadavek, tudíž se může stát, že budou navraceny i mnohé nerelevantní informace [4].

Obě metody mají své výhody i nevýhody a nelze obecně říct, která z nich je lepší. Na základě vlastností však lze usoudit, že *303 URI* jsou vhodnější pro velké zdroje dat, kde je potřeba objekty důsledně rozlišovat a vracet maximálně relevantní informace, jelikož objem souvisejících dat je obrovský. Zatímco *Hash URI* se používají například pro definici slovníku, jelikož je často užitečnější získat rovnou dokument s popisem celého slovníku. V praxi se většinou používá kombinace obou metod [4].

2. Hypertext Transfer Protocol

Základem WWW je HyperText Transfer Protocol (HTTP), který definuje pravidla komunikace mezi klientem a serverem. Jedná se o bezstavový protokol. To znamená, že server si neudrhuje žádné informace o svých klientech. Obdrží-li dotaz, odpoví na něj a tím pro něj skončila jedna ucelená HTTP transakce. Pokud zanedlouho dostane dotaz od téhož klienta, nedává si jej do žádných souvislostí s dotazem předchozím (přestože byl třeba vyvolán některým z odkazů na stránce, kterou tomuto klientovi před chvílkou odeslal). Veškeré stavové informace (např. adresu aktuální stránky) si musí pamatovat klient [7].

3. Resource Description Framework

RDF klade důraz na jednoduchost automatického zpracování webových zdrojů, a proto je základním kamenem sémantického webu. Strukturu tohoto frameworku zajišťuje jazyk XML (této reprezentaci se dává zkratka RDF/XML) [8].

RDF je založeno na obecné myšlence, že jakékoliv tvrzení se dá rozložit na tři části – subjekt, predikát a objekt. Těmto třem částem se říká RDF trojice (triple). Subjekt představuje entitu, která je touto trojicí popisována. Predikát vyjadřuje určitý vztah či vlastnost a objekt je hodnotou predikátu. Pro ukázkou, tvrzení „Josef Novák bydlí v Praze“ by bylo v RDF reprezentaci rozloženo na tři části:

- *subjekt* - Josef Novák
- *predikát* - bydlí
- *objekt* – Praha

V praxi je subjekt i predikát reprezentován vždy pomocí URI.

Pro Josefa Nováka, aby objekt reálného světa, by tedy mělo existovat unikátní URI identifikující jeho osobu, řekněme například <http://example.org/people/JosefNovak> [4].

Výběr a použití slovníků pro popis dat

RDF poskytuje obecný, abstraktní datový model pro popis. Nicméně, to neposkytuje žádné konkrétní domény pro popis třídy věcí na světě. Tato funkce je používána pomocí slovníků taxonomie a ontologie, vyjádřené ve SKOS (Simple Knowledge Organization System), RDFS (RDF Schema) a OWL (Web Ontology Language) [1].

SKOS

SKOS je slovník pro vyjádření konceptuálních hierarchií, často označovaných jako taxonomie. Zatímco RDFS a OWL poskytují slovní zásobu pro popis koncepčních modelů z hlediska tříd a jejich vlastností.

RDFS

RDFS je jazyk pro popis ontologií v RDF, který často odkazuje na slovníky. RDFS poskytuje základní mechanismus pro popis tříd a jejich vlastností.

Základními konstrukty RDFS jsou:

- `rdfs:Class` – třída pro zdroje (subjekty, objekty)
- `rdfs:Property` – třída pro vlastnosti (predikáty)
- `rdf:type` – vlastnost zdroje definující jeho typ - příslušnost ke třídě

OWL

OWL doplňuje RDFS o další konstrukty umožňující podrobnější a přesnější popis cílové domény.

4. Protocol and RDF Query Language

SPARQL je dotazovacím jazykem nad RDF daty. Jde o nástroj, který pro svět grafových databází představuje to, co SQL pro relační databáze. Je to jedna z klíčových technologií sémantického webu, jež se v roce 2008 stala oficiálně W3C doporučením. Dotazovací jazyk SPARQL je založen na specifikaci množiny vzoru trojic, která představuje základní vzor grafu. Tyto vzory trojic jsou obdobou RDF trojic s tím, že všechny tři části (subjekt, predikát, objekt) mohou být v rámci dotazu proměnnou [4].

Příklad jednoduchého SPARQL dotazu, který by vrátil e-mailové adresy všech Josefů Nováků žijících v Praze [4].

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?email
WHERE {
  ?person foaf:name "Josef Novák";
          foaf:base_near http://dbpedia.org/resource/Prague;
          foaf:mbox ?email.
}
```

1.4 Vstupní data pro publikování linked dat

Prvořadým hlediskem při publikování linked dat jsou vstupní data.

1. Statická strukturovaná data

Statická vstupní data se mohou skládat např. z CSV souborů, tabulek aplikace Excel nebo XML databázových souborů. Tato data musí podstoupit proces přeměny, který vyprodukuje statické RDF soubory, nebo jsou data převedena přímo do úložiště RDF. Pro tuto konverzi se používají tzv. RDFizing nástroje, kde jsou statické soubory již ve formě RDF a linked data mohou být jednoduše publikována na webu pomocí webového serveru [5].

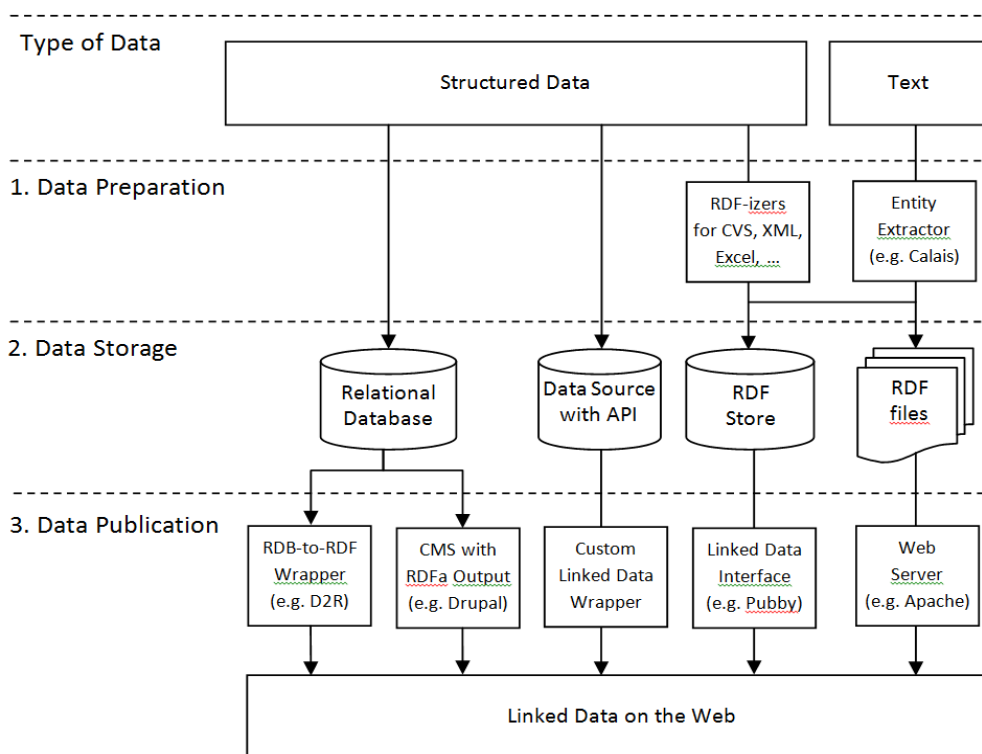
2. Textové dokumenty

V případě, že se vstupní data skládají z textových dokumentů v přirozeném jazyce, např. série novinových nebo obchodních zpráv, je možné předat tyto dokumenty prostřednictvím např. technologie Ontos nebo DBPedia Spotlight.

Ontos je poskytovatel sémantických technologií. Softwarové řešení pracuje na základě ontologie a zpracování přirozeného jazyka. S pomocí produktů Ontos, lze snadno ovládat, popsat a zhodnotit komplexní data z dokumentů, e-mailů či webových stránek. Odkaz na domovské stránky <http://www.ontos.com/>.

DBpedia Spotlight poskytuje řešení pro propojení nestrukturovaných informačních zdrojů. Odkaz na domovské stránky <https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki>.

Nejčastější vzory publikování dat zobrazuje Obr. 1.



Obr. 1: Publikování linked dat [5].

Jaký je doporučen objem dat?

Množství dat, které chceme publikovat, bude mít silný vliv na vzor, který si vybereme. Pokud chceme publikovat jen malé množství dat, např. jen několik RDF trojic o jednom subjektu, pak je žádoucí, aby byl použit statický soubor RDF. Může to vyžadovat větší úsilí na ruční správu dat v případě, že musí být udržována stejná data v jiném umístění nebo formátu. Na druhou stranu se vyhýbá větším instalačním nákladům, spojeným s technicky složitými vzory [5].

Jak často data měnit?

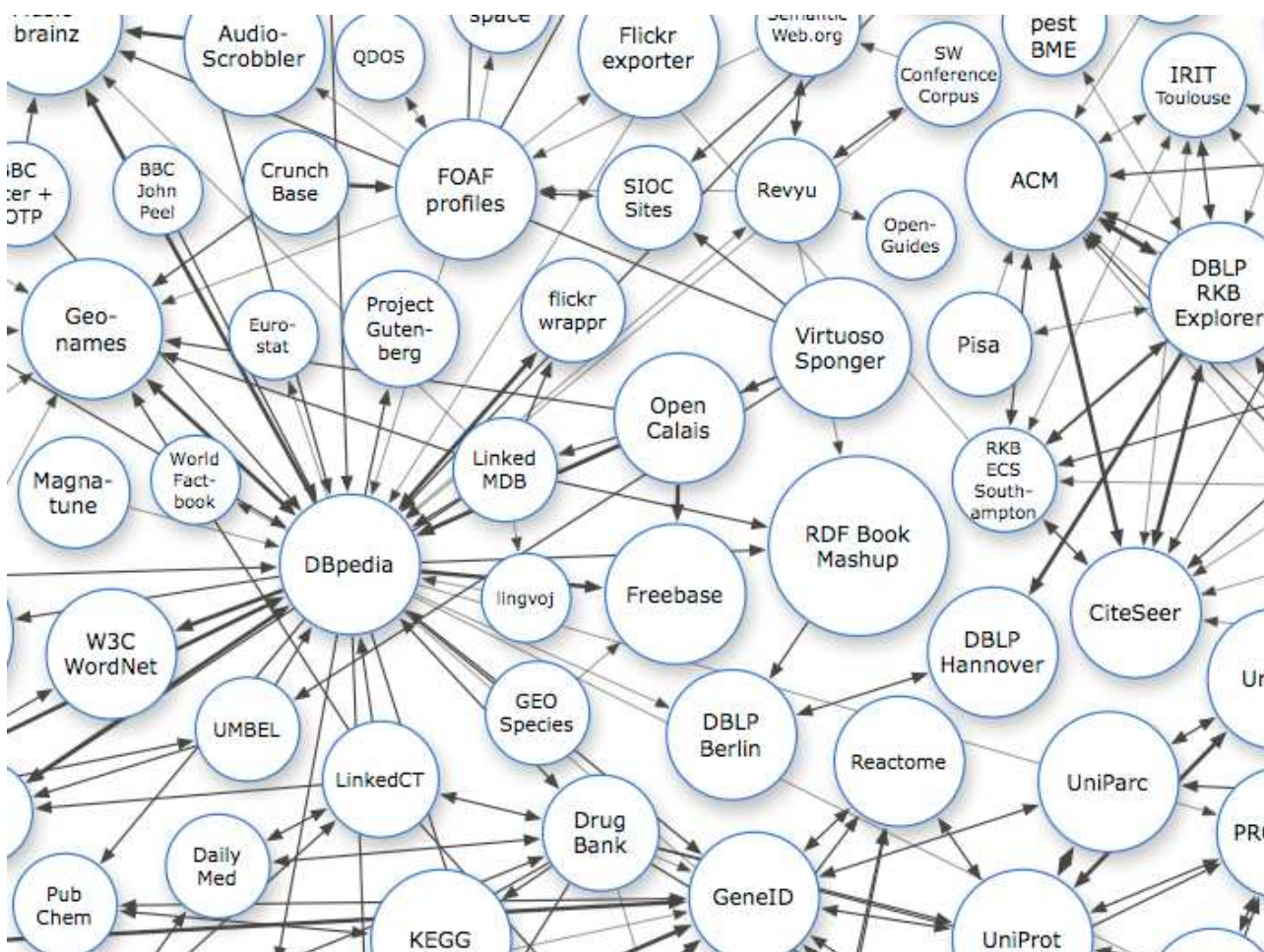
Rychlost změny s obyčejným objemem dat bude v datovém souboru ovlivněna opět vzorem, který si vybereme. Pro data, která se mění jen zřídka (např. historické záznamy), může být vhodná publikace pomocí statických souborů. Pokud se data mění často, bude vhodnější použít úložiště, které umožňuje časté změny [5].

1.5 Web of data - linked data

Výsledkem publikování linked dat na webu je vytvoření systému strukturovaných dat, která jsou vzájemně provázána odkazy. Tento systém se označuje jako tzv. Web of data. Web of data tvoří obří globální graf [3].

Obsah cloudů (viz. Obr. 2) je různorodý. Zahrnuje údaje o zeměpisných místech, lidech, firmách, knihách, vědeckých publikacích, filmech, hudbě, televizních a rozhlasových programech, lécích, on-line komunitách, statistických údajích, výsledcích sčítání lidu, recenzích a dalších [1].

Web of data je obecný a může obsahovat jakýkoli typ dat. Každý může publikovat data na webu. Část datových sad, které jsou publikovány pomocí linked dat, jsou zobrazeny na Obr. 2.



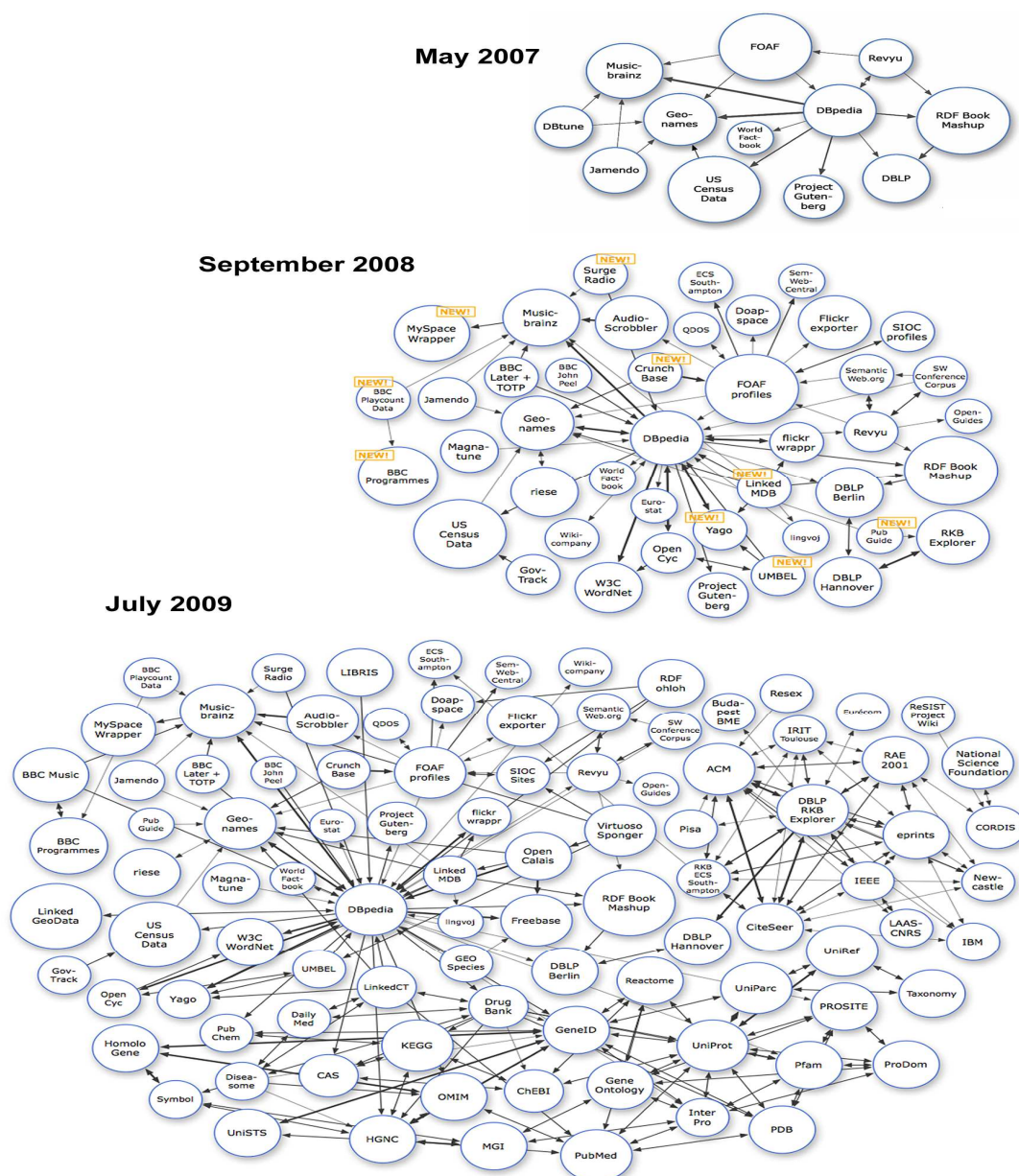
Obr. 2: Část LOD cloud diagramu. Odkaz na celý cloud diagram: <http://lod-cloud.net/>.

Počátky Web of data spočívají v úsilí komunity sémantického webu a zejména v aktivitách **W3C Linking Open Data (LOD)**. Cílem založení projektu bylo pomocí identifikace stávajících údajů, které jsou k dispozici v rámci otevřených licencí, převést údaje do RDF podle linked dat principů. Projekt byl vždy otevřený pro toho, kdo zveřejňuje údaje v souladu s principy linked dat. Každý

uzel v diagramu představuje soubor dat publikován jako linked data. Všechna data, která jsou zveřejněna na webu, v souladu s principy linked dat, se stávají součástí jednoho globálního datového prostoru. Obecně platí, že aplikace jsou postavené na využití následujících vlastností linked dat architektury:

1. Standardizovaná reprezentace dat a přístup k datům.
2. Otevřenost webových dat.

Související datové architektury jsou otevřené a umožňují objev nových datových zdrojů za běhu. To umožňuje aplikacím automaticky využít nové datové zdroje, jakmile jsou k dispozici na Web of data.



Obr. 3: Ukazuje, jak počet linked dat na webu jako vzrostl [5].

1.6 Topologie datových sad

Datové soubory jsou rozděleny do následujících tematických oblastí: geografie, vláda, média, knihovny, přírodní vědy, prodej a obchod, obsah vytvářený uživatelem a doménami datových sad [5].

Vybrané datové sady:

Geografická data

Geografie je faktorem, který může často propojovat informace z různých tematických oblastí. Je to patrné ve Web of data, kde např. Geonames [11] slouží jako rozbočovač pro jiné datové sady, které mají nějaký zeměpisný prvek. Geonames je geografická databáze s otevřenou licencí, která nabízí linked data o 8 milionech míst.

Druhá významná data uvedená v této oblasti jsou LinkedGeoData. Využívají údaje shromážděné v rámci projektu OpenStreetMap, který poskytuje informace o více než 350 milionech prostorových vlastnostech. Kdykoli je možné, je umístění v Geonames a LinkedGeoData propojeno s odpovídajícími místy v DBPedia, která je jádrem linked dat o zeměpisných oblastech. DBPedia je komunita, která usiluje o získávání strukturovaných informací z Wikipedie [12]. Více o DBPedia na <http://dbpedia.org/About>.

Media Data

Jedna z prvních velkých organizací, která rozpoznala potenciál linked dat a přijala zásady a technologie do jejich pracovních postupů pro správu obsahu byl British Broadcasting Corporation (BBC). Webové stránky nazvané - /music (<http://www.bbc.co.uk/music>), vydávají linked data o každém umělci, jehož hudba se hrála na rozhlasových stanicích BBC. Každý z nich byl v pozadí označen prostřednictvím HTTP URI a byl popsán v RDF [13].

Vládní data

Státní orgány a organizace veřejného sektoru produkují velké množství dat, od ekonomických statistik, registrů podniků a vlastnictví půdy, zprávy o výkonu škol, statistiky trestné činnosti až po údaje o průběhu hlasování volených zástupců. Tvorba takto snadno přístupných dat umožňuje organizacím a členům veřejnosti práci s daty - analyzovat a objevovat nové poznatky, vytvářet nástroje, které sdělí tyto poznatky, a tím pomohou občanům přijímat informovaná rozhodnutí [14].

Knihovny a vzdělávání

Významným vývojem je integrace katalogů knihoven v celosvětovém měřítku a propojení obsahu několika katalogů knihoven, např. podle tématu, místa nebo historického období. Propojují se katalogy knihoven s informací od třetích stran (obrázek a video archivy, nebo znalostní báze, jako DBPedia). Příkladem jsou the American Library of Congress a German National Library of Economics, které zveřejnily své předmětové heslo taxonomie jako linked data.

Nedávno byly linked data principy a technologie přijaty hlavními uživateli sociálních sítí. Nejvýznamnějším příkladem je vývoj a přijetí Facebooku, který využívá Open Graph Protocol. Open Graph Protocol umožňuje webovým uživatelům vyjádřit několik základních informací o věcech popsaných na svých webových stránkách pomocí RDFa (Technologie pro přenos strukturovaných informací uvnitř webových stránek. RDFa je jeden ze způsobů zápisu datového formátu RDF. K přenosu dat v RDF používá atributy XHTML nebo HTML elementů webové stránky).

Facebooku to umožňuje snadněji sdílet data z webových stránek po celém webu. Hlavním úkolem pro Open Graph Protocol je umožnit větší míru propojení mezi zdroji dat [5]. Více o Open Graph Protocol na <http://ogp.me/>.

1.7 Metadata

Aby se zvýšila užitečnost pro uživatele, měla by být linked data zveřejněna spolu s několika typy metadat. Cílem je umožnit klientům zhodnotit kvalitu publikovaných dat a určit, zda jim chtějí důvěřovat. Každý údaj by měl být doprovázen tzv. Meta-informacemi. Jsou to informace o tvůrci či poskytovateli, datu vytvoření, způsobu vytvoření a licenčních podmínkách [15].

Metadata dávají uživateli jasnost o původu a aktuálnosti datových souborů a podmínky, za kterých mohou být data opakovaně použita. Kromě toho je popis souboru například ukazatelem na zdroje v souboru dat, čímž přispívá k lepšímu vyhledávání a indexování dat.

K dispozici jsou dva hlavní mechanismy pro publikování popisu souboru dat: Semantic Sitemaps [16] a void descriptions [17].

1.8 Linked data aplikace

Vzhledem k tomu, že dostupnost linked dat je poměrně novým jevem, prezentované aplikace jsou většinou první generace aplikací, které podstupují významnou evoluci.

Linked data aplikace mohou být rozděleny do dvou kategorií: *obecné aplikace* a *aplikace pro specifické domény*.

1. Obecné aplikace

Obecné aplikace mohou zpracovávat data z libovolné aktuální domény. Existují dva základní typy obecných datových aplikací: **Linked Data browsers** a **Linked Data search engines**.

Linked Data Browsers

Linked data prohlížeče umožňují uživatelům přecházet mezi zdroji dat pomocí odkazů vyjádřených jako RDF trojice. Mezi datové prohlížeče, které sledují původ dat při sloučení údajů o téže věci z různých zdrojů, patří např. **Marbles**. Je to datový prohlížeč, který sleduje původ dat při sloučení údajů z různých zdrojů.

Na *Obr. 4*, prohlížeč Marbles zobrazuje údaje o Tim Berners-Lee, které byly sloučeny z různých zdrojů. Barevné tečky označují datové zdroje, ze kterých byly údaje sloučeny [5].

Tim Berners-Lee

<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>

- Person
- <http://www.w3.org/2000/10/swap/pim/contact#Male>


label

- Tim Berners-Lee

sameAs

- [Tim Berners-Lee \(also at www4.wiss.fu-berlin.de\)](#)

image



Webinks

<http://www.w3.org/People/Berners-Lee/>

name

- Tim Berners-Lee
- Timothy Berners-Lee
- Tim Berners Lee

Given name

- Timothy

family name

- Berners-Lee

sha1sum of a personal mailbox URI name

- 985c47c5a70db7407210cef0e4e8f5374a525c5c

workplace homepage

- <http://www.w3.org/>

Obr. 4: Prohlížeč Marbles [5].

Linked Data search engines

V tradičním hypertextovém webu je prohlížení a vyhledávání často viděno jako dvoudominantní způsob interakce. Zatímco prohlížeče poskytují mechanismy pro navigaci v informačním prostoru, vyhledávače jsou často místem, v němž tento proces navigace začíná. Řada vyhledávačů byla vyvinuta procházením linked dat z webu pomocí odkazů RDF. Obecně řečeno, tyto služby mohou být rozděleny do dvou kategorií: **Human-oriented Search Engines** a **Application-oriented Indexes**. Human-oriented Search Engines je např. vyhledávač Falcons a Application-oriented Indexes je např. Sindice [5].

Falcons (viz. Obr. 5) poskytuje uživatelům možnost vyhledávání objektů a dokumentů, z nichž každý vede k mírně odlišným výsledkům. Hledání objektů na Obr. 4, je vhodné pro vyhledávání lidí, míst a dalších. Zatímco funkce pro vyhledávání dokumentů poskytuje více tradiční vyhledávače, kdy výsledky přejdou na RDF dokumenty, které obsahují zadané hledané výrazy [5].

The screenshot shows the Falcons search engine interface. At the top, there is a search bar with the text 'Berlin' and a 'Search Objects' button. Below the search bar, there is a section titled 'Specify a type:' with a grid of categories: Agent, Event, Motion Picture Film, Album, Facility, Organization (highlighted in blue), Building, Group (highlighted in blue), Person, City, Landmark, State, Concept, Location (highlighted in blue), and Subject. Below this, a status bar indicates 'Objects 1 - 10 of 42,186 for your search Berlin (2.4 seconds)'. The main content area displays three search results for 'Berlin':

- Berlin** is a *State, Capital, City*
 - abstract: :Berlin redirige para aqui. Para outros significados, v... - [From dbpedia.org »](#)
 - has subject: Category:13th_century_establishments - [From dbpedia.org »](#)
 - hasPhotoCollection: Berlin - [From dbpedia.org »](#)
 - <http://dbpedia.org/resource/Berlin> - Described in 1855 documents
- _Berlin** is a *Thing, _Category-3AStadt*
 - hasArticle: Berlin - [From www.sembase.at »](#)
 - isDefinedBy: <http://www.sembase.at/index.php/Special:ExportRDF/Berlin> - [From www.sembase.at »](#)
 - label: Berlin - [From www.sembase.at »](#)
 - http://wiki.sembase.at/index.php/_Berlin - Described in 17 documents
- Berlin** is a *Thing, Subject, City*
 - isDefinedBy: <http://ontoworld.org/wiki/Special:ExportRDF/Berlin> - [From ontoworld.org »](#)
 - page: Berlin - [From ontoworld.org »](#)
 - label: Berlin - [From ontoworld.org »](#)
 - <http://ontoworld.org/wiki/Special:URIResolver/Berlin> - Described in 16 documents

Obr. 5: Falcons - výsledky vyhledávání na klíčové slovo "Berlin" [5].

Sindice shromažďuje webová data v mnoha ohledech a nabízí vyhledávání a dotazování v rámci těchto údajů. Aktualizace dat probíhá po pár minutách. K dispozici jsou specializované API a nástroje [5].

2. Aplikace pro specifické domény

Zatímco Linked Data Browsers a Linked Data search engines popsané výše poskytují do značné míry obecné funkce, řada služeb, která byla vyvinuta, nabízí specifické funkce. Existují různé vzájemně propojené datové aplikace, které splňují potřeby specifických uživatelů jako je např. DBpedia mobile.

DBpedia Mobile

Aplikace DBpedia Mobile je orientována na turistickou prohlídku města. Prohlížeč je určený ke spuštění na iPhonech nebo jiném mobilním zařízení. Aplikace na základě aktuální polohy GPS z mobilního zařízení, poskytuje umístění okolních lokalit z DBpedia. Obr. 6 ukazuje zobrazování dat z DBpedia o Brandenburské bráně v Berlíně. Kromě přístupu k webovým datům, DBpedia Mobile umožňuje uživatelům publikovat své aktuální umístění, obrázky a recenze na web, jako linked data, takže mohou být použity v jiných aplikacích [1].



Obr. 6: Aplikace DBpedia mobile [1].

1.9 Ochrana osobních údajů

Konečným cílem linked dat je použití globální databáze na webu. Realizace této vize byla přínosem v mnoha oblastech, přináší sebou však také nebezpečí u druhých. Jednou problematickou oblastí jsou možnosti porušování soukromí, které vyplývají z integrace dat z různých zdrojů.

Ochrana soukromí v kontextu linked dat bude pravděpodobně vyžadovat kombinaci technických a právních prostředků, spolu s větší informovaností uživatelů o tom, jaké údaje mohou poskytovat a v jakém kontextu [5].

1.10 Příklady konferencí a vybrané workshopy na téma linked data

Svou cestu si název linked data našel jak do názvů konferencí (např. Dublin Core 2009: Semantic Interoperability of Linked Data), tak do příspěvků z periodik, které se zabývají problematikou knihovnictví (např. Library Journal nebo Computers in Libraries) [3].

- Linked Data on the Web (LDOW2012) Workshop at WWW2012
Odkaz: <http://events.linkedata.org/ldow2012/>
- Linked Data on the Web (LDOW2011) Workshop at WWW2011
Odkaz: <http://events.linkedata.org/ldow2011/>
- Linked Data on the Web (LDOW2010) Workshop at WWW2010
Odkaz: <http://events.linkedata.org/ldow2010/>
- 2nd International Workshop on Consuming Linked Data (COLD 2011) at ISWC 2011
Odkaz: <http://km.aifb.kit.edu/ws/cold2011/>
- 1st International Workshop on Consuming Linked Data (COLD 2010) at ISWC 2010
Odkaz: <http://people.aifb.kit.edu/aha/2010/cold/>
- Workshop On Linked Spatiotemporal Data (LSTD2010) at GIScience 2010
Odkaz: <http://stko.geog.ucsb.edu/lstd2010/>

1.11 Zajímavé video popisující linked data

Linked Open Data - What is it? - Europeana - think culture

Odkaz na film: <http://vimeo.com/36752317>

2 OPEN DATA

2.1 Co jsou open data?

“A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike” [29].

Otevřená data jsou dostupná, bezplatná data na internetu, ve strukturované a strojově čitelné podobě. Otevřená data používají formát dat, jehož specifikace je volně dostupná, a právní podmínky neomezují nikoho v jejich použití a volném zpracování, a to i v rámci softwarových aplikací. Tyto údaje mohou pocházet z univerzit, nevládních organizací, soukromých firem nebo veřejné správy [18].

Co je ale přesně myšleno tou otevřenou podobou?

- 1. Technická otevřenost**, tj. zveřejnění dat ve standardním strojově čitelném formátu.
- 2. Legislativní otevřenost**, tj. zveřejnění dat pod otevřenou licenci.
- 3. Dostupnost a původ**, tj. datové sady jsou zveřejňovány jako jeden celek a jsou nezměnné.
- 4. Přehlednost**, tj. katalogizace datových sad v katalogu dat pro usnadnění vyhledávání.

Myšlenkou open dat je, že některé údaje by měly být pro každého volně k dispozici. Publikování dat by mělo být bez omezení patentů autorských práv nebo jiných kontrolních mechanismů.

Cíle open dat jsou podobné jako cíle open source. Termín open data získal na popularitě hlavně s nástupem internetu. Zejména s uvedením open dat vládní iniciativy, jako jsou např. Data.gov. Účelem Data.gov je zlepšit přístup veřejnosti k strojově čitelným datovým souborům, vytvořených výkonnou mocí federální vlády. Data.gov zvyšuje schopnost veřejnosti snadno najít, stáhnout a používat datové sady. Poskytuje popisy federálních souborů dat (metadat), informace o tom, jak získat přístup k datovým souborům, a nástroje, které využívají vládní datové sady [19].

Více na stránkách (<http://www.data.gov/>).

Dalším cílem open dat je motivovat vlády a organizace, aby byly informace volně dostupné a snadno přístupné on-line. Výhody open dat jsou ekonomické, prostřednictvím identifikace nových obchodních příležitostí a sociální, prostřednictvím zvýšené transparentnosti, účasti a odpovědnosti [20].

Vlády tradičně ochraňují svá data a jako důvody, uvádějí národní bezpečnost a soukromí občanů. Nicméně otevřená vládní data se netýkají citlivých informací, ale základních veřejných údajů týkající se dopravy, infrastruktury, vzdělávání, zdravotnictví, kriminality, životního prostředí, atd. Mnoho lidí věří, že veřejné údaje by měly být open data, protože se jedná o informace týkající se veřejnosti, a také proto, že byly shromážděny pomocí veřejných financí, tj. daní [21].

Open data se často zaměřují na netextový materiál, jako jsou mapy, genomy, matematické a vědecké vzorce, medicínská data a biologické rozmanitosti. Vznikají zde často problémy, protože data jsou komerčně cenná. Přístup nebo znovu-použití dat je kontrolováno organizacemi, jak veřejnými tak soukromými. Kontrola může být provedena prostřednictvím přístupových omezení, licencí, autorských práv, patentů a poplatků za přístup nebo opětovného použití.

Zastánci open dat tvrdí, že tato omezení jsou proti společnému dobru, a že by tyto údaje měly být zpřístupněny bez omezení a bez poplatků [20].

Tvůrci dat často nepovažují za potřebné uvést podmínky vlastnictví, udělování licencí a opětovného využití. Například, vědci nepozorují publikování údajů vyplývajících z jejich prací.

2.2 Otevřené formáty souborů

V otevřených formátech souborů jsou zveřejněné informace. Jinými slovy, digitální báze, ve které jsou informace uloženy - může být "otevřená" nebo "uzavřená".

Otevřený formát dat je ten, který je bezplatně k dispozici každému a obsahuje specifikaci pro určitý software. Pokud jsou formáty souboru "uzavřené", může to být buď proto, že specifikace není veřejně dostupná, nebo proto, že specifikace byla zveřejněna a opětovné použití je omezené [22].

XML

Extensible Markup Language je obecný značkovací jazyk, který byl vyvinut a standardizován konsorciem W3C. Je zjednodušenou podobou staršího jazyka SGML. Umožňuje snadné vytváření konkrétních značkovacích jazyků (tzv. aplikací) pro různé účely a různé typy dat. Používá se také pro serializaci dat. Zpracování XML je podporováno řadou nástrojů a programovacích jazyků. Jazyk je určen především pro výměnu dat mezi aplikacemi a pro publikování dokumentů, u kterých popisuje strukturu z hlediska věcného obsahu jednotlivých částí. Prezentace dokumentu (vzhled) může být definována pomocí kaskádových stylů. Při používání XML dokumentu potřebujeme také dokument zobrazit. XML samo o sobě žádné prostředky pro definici vzhledu nenabízí. Existuje ale několik stylových jazyků, které umožňují definovat, jak se mají jednotlivé elementy zobrazit. Souboru pravidel nebo příkazů, které definují, jak se dokument převede do jiného formátu, se říká

styl. Další možností zpracování je transformace do jiného typu dokumentu, nebo do jiné aplikace XML. Specifikace XML konsorcia W3C je zdarma přístupná všem. Každý tak může bez problémů do svých aplikací implementovat podporu XML [30].

JSON

JavaScript Object Notation je odlehčený formát pro výměnu dat. Je jednoduše čitelný i zapisovatelný člověkem a snadno analyzovatelný i generovatelný strojově. Je založen na podmnožině programovacího jazyka JavaScript. JSON je textový, na jazyce zcela nezávislý formát. Využívá však konvence dobře známé programátorům jazyků rodiny C (C, C++, C#, Java, JavaScript, Perl, Python a dalších). Díky tomu je formát JSON ideálním pro výměnu dat [31].

RDF

Viz kapitola 1.3 – Resource Description Framework.

Textové dokumenty

Standardem pro textové soubory jsou následující formáty:

Rich Text Format (.rtf) – Formát na ukládání a výměnu textových dokumentů, který byl vyvinutý společností Microsoft. Pro formát RTF je dlouhodobě veřejně vydávána specifikace, čímž se zásadně liší od uzavřeného formátu DOC. Většina kancelářských programů na zpracování textu má implementovanou alespoň základní podporu formátu RTF. Tento formát není určený a ani vhodný pro dokumenty s obrázky. Hodně programů pro práci s formátem RTF totiž vložené obrázky ukládá jako bitmapy bez komprese, čímž velikost souborů velmi narůstá (až na desítky megabajtů). Tak velké soubory nejsou vhodné pro zveřejňování dokumentů na internetu. Některé formáty obrázků navíc nejsou podporované ve všech programech pro práci s RTF a tudíž se nemusí vůbec zobrazit. Zde je vhodnější použít formát PDF [30].

Hypertext Markup Language (.html, .htm) - Hypertextový značkový jazyk. Předností formátu HTML oproti jiným textovým formátům, je možnost čtení HTML souborů přímo v libovolném internetovém prohlížeči (při dodržování W3C standardů). Některé programy, které přečtou formát HTML a jsou zdarma: Firefox, Google Chrome, Opera.

Portable Document Format (.pdf) - Formát na ukládání a výměnu textových dokumentů, které jsou primárně určené pro čtení. Velkou předností formátu PDF oproti jiným formátům je to, že jako jediný dokáže plně zachovávat vzhled a formátování dokumentu v různých prohlížečích, díky čemuž je to nejčastěji používaný formát pro finální elektronické dokumenty, tiskoviny apod. [30].

Open Document Format (.odt) - ODT je formát, který se snaží nahradit některé z dnes používaných formátů, zatím je však poměrně málo rozšířený. Formát je možné použít i pro dokumenty, u kterých se předpokládá další úprava. Některé programy, které přečtou a upravují formát odt a jsou zdarma: Open Office, AbiWord, Google Docs, KWord, Lotus Symphony [30].

Text Format (.txt) - Výhodou formátu TXT je možnost jeho čtení přímo v internetových prohlížečích. S tímto formátem dokáže pracovat velké množství programů, mimo jiné třeba i souborové manažery (TotalComander), prohlížeče fotek (XnView) apod. Dá se říci, že tento formát přečtete prakticky v libovolném softwaru. Tento formát nelze doplňovat obrázky. Některé programy, které přečtou a upravují formát TXT a jsou zdarma: gedit, Notepad++ [30].

CSV

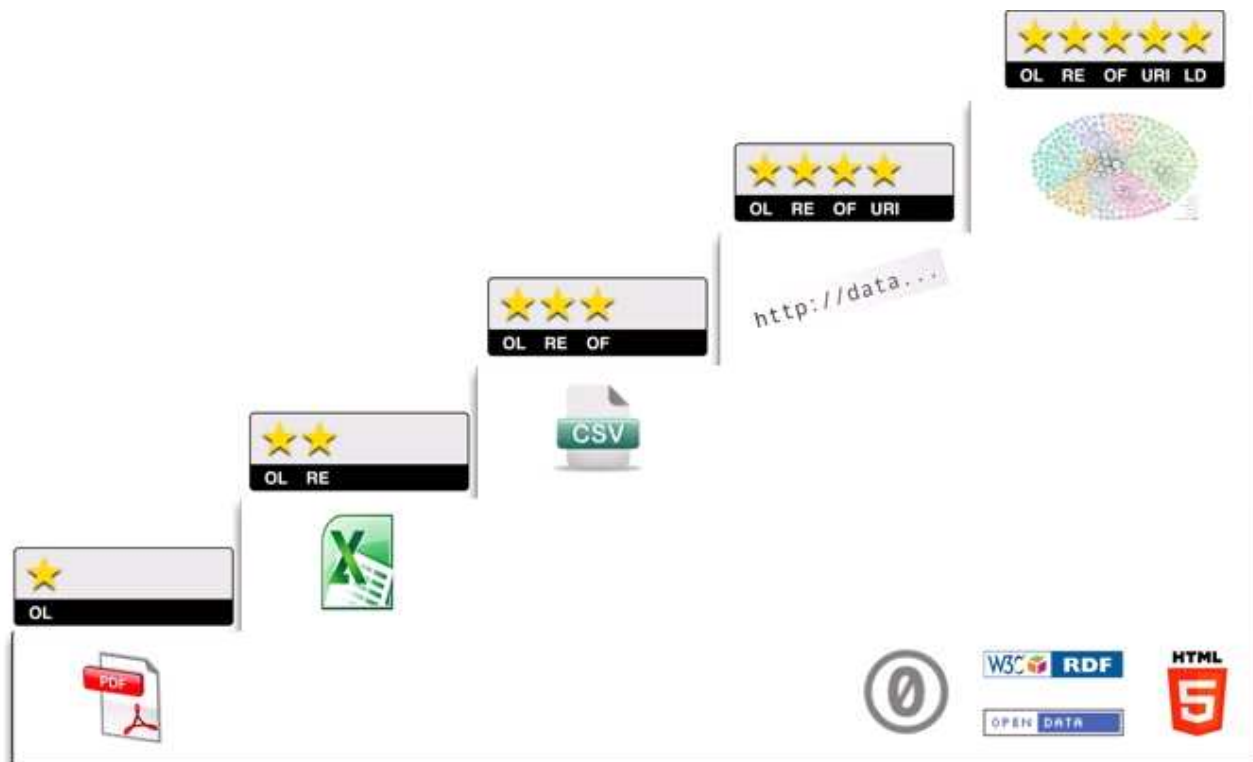
Comma-separated values je jednoduchý souborový formát určený pro výměnu tabulkových dat. Soubor ve formátu CSV sestává z řádků, ve kterých jsou jednotlivé položky odděleny znakem (,). Jelikož se v některých jazycích včetně češtiny čárka používá v číslech jako oddělovač desetinných míst, existují varianty, které používají jiný znak pro oddělování položek než čárku, nejčastěji středník. Variantu se středníkem (ale stále pod názvem CSV) používá např. Microsoft Excel v české verzi Microsoft Windows (řídí se oddělovačem zadaným v Místním a jazykovém nastavení). Díky jednoduchosti, nenáročnosti a čitelnosti i bez specializovaného software se tento formát používá pro výměnu informací mezi různými systémy. Soubory CSV mohou být velmi užitečné, protože je jejich formát kompaktní, a tedy vhodný pro přenos velkých souborů dat se stejnou strukturou [32].

2.3 Publikování open dat

Tim Berners-Lee navrhl 5-hvězdičkový systém pro open data.

5 ★ Open Data

- ★ publikování dat na webu
- ★★ publikování dat ve strojově zpracovatelné podobě – strukturovaná data
- ★★★ používat standardní formáty (např. CSV místo v Excelu)
- ★★★★ publikovat data pod otevřenou licenci
- ★★★★★ seznam dat v katalogu dat



Obr. 7: 5-hvězdičkový systém pro open data <http://5stardata.info/>.



Obr. 8: 5-hvězdičkový systém pro open data <http://www.w3.org/DesignIssues/LinkedData.html>.

Příklady otevřených datových technologií

CKAN

Comprehensive Knowledge Archive Network je open-source datový portál iniciovaný Open Knowledge Foundation. CKAN usnadňuje publikování, sdílení a vyhledání dat tím, že poskytuje výkonné databáze pro katalogizaci a ukládání datových souborů. CKAN je používán ve Velké Británii, norskými a nizozemskými vládami, místní správou a specializovanými datovými vydavateli [21].

INSPIRE

Je velmi důležité, aby do iniciativy open dat, jako veřejných informací, byla zahrnuta i prostorová data. Směrnice INSPIRE vstoupila v platnost dne 15. května 2007 a bude realizována v různých fázích. Plná realizace je plánována do roku 2019. Směrnice INSPIRE je zaměřena na vytvoření infrastruktury prostorových dat. To umožní sdílení prostorových informací o životním prostředí mezi organizacemi veřejného sektoru. Tím se usnadní přístup veřejnosti k prostorovým informacím v celé Evropě [21].

2.4 Příklady užití vládních open dat v USA

Vládní open data, jsou obrovský zdroj, který je dosud z velké části nevyužitý. Mnoho jednotlivců a organizací shromažďuje širokou škálu různých typů dat. Vláda je v této souvislosti zvláště významná. Vládní údaje jsou podle zákona veřejnými údaji, a proto by měli být otevřeny a zpřístupněny pro ostatní uživatele. Proč je o ně takový zájem? Existuje mnoho různých skupin lidí a organizací, kteří mohou mít prospěch z dostupnosti open dat, včetně samotné vlády [23].

Otevřené vládní iniciativy přinášejí nové bohatství veřejných dat na webu. Umožňují občanům a vládě sdílet společný obraz inteligence, který řídí rozhodování v celé zemi. Skutečná hodnota otevření vládních dat spočívá v tom, jak ji používají orgány a veřejnost. Poskytování přístupu k datům je nezbytným prvním krokem, ale sdílení v informativní podobě je klíčem k podpoře nové úrovně vládní komunikace a efektivity. GIS technologie odemykají potenciál otevřených dat tím, že je dávají do prostorového kontextu, který umožňuje jednotlivcům a vládě, aby lépe pochopili jejich světovou informativnost.

Po celá desetiletí se americká vláda opírala o společnost ESRI. O GIS platformy pro podporu úlohy kritického a každodenní provozu. Nyní využívá vláda a soukromý sektor GIS jako zdroj map, a sdílí

obrovské množství nově dostupných dat, a to je zásadní krok v usnadnění vládních open dat [23].

Se zahájením Data.gov je k dispozici několik on-line autoritativních vládních datových souborů. Vláda v USA investovala do vytváření a udržování geografických informací. Zahrnuty jsou výsledky sčítání lidu, geologické průzkumy, zachování a využití půdních informací, hodnocení infrastruktury, daňové evidence, zdravotní statistiky a údaje o vzdělávání. On-line mapové aplikace se jeví jako jeden z nejúčinnějších prostředků pro komunikaci vládních údajů.

Sdílením informací, může vláda jasněji vyjádřit své záměry a posílit informovanost a sledovat reakci veřejnosti. Technologie ArcGIS poskytuje platformu pro otevřenou vládní iniciativu, protože podporuje poskytování mapových služeb a aplikací napříč webem. Tato technologie poskytuje základ pro využití geoprostorových investic a efektivní správu dat, provádění prostorových analýz a nasazení mobilních služeb [23].

2.5 Veřejná správa jako zdroj dat

Málokdy, jsou jako zdroje dat právě aplikace veřejné správy. Ty přitom produkují mnoho různých dat, které jsou pro občany mimořádně zajímavé. Pokud by k datům existovaly smysluplné online aplikace, občané by je využívali stejně, jako dnes využívají aplikace k nakupování nebo vyhledávání. Kouzlo je v tom, že veřejná správa (úřad, město, ...) zveřejní svá data na internetu v otevřené podobě tak, aby s nimi mohla volně pracovat veřejnost. Široká veřejnost samozřejmě nechce pracovat přímo s daty. Vyžaduje zajímavé aplikace, které s daty pracují. Důležité je, že data jsou dostupná komukoliv a nikdo není diskriminován. Jak ukazují příklady ze světa, není třeba se obávat nedostatku aplikací, které jsou veřejnosti dostupné zdarma [24].

Zveřejnění dat v otevřené podobě je přitom pro veřejnou správu velice levné, řádově levnější než vytváření a údržba aplikací. Data už totiž veřejná správa dávno má. Buď je má uloženy v databázích svých informačních systémů, nebo leží v podobě „excelovských“ tabulek v počítačích úředníků. V některých oblastech je již splněna technická otevřenost. Např. informační systém Ministerstva financí ARES (Administrativní registr ekonomických subjektů) či ÚFIS (Prezentační systém finančních a účetních informací státu) umožňují technicky otevřený přístup k datům ve svých databázích. V některých oblastech samozřejmě nelze k otevřenému zveřejnění dat přistoupit (např. strategická data o rozvodech elektrické sítě nebo údaje chráněné zákonem o ochraně osobních údajů). To je ale jen malá část dat, která veřejná správa vytváří. Existuje mnoho dat, která zveřejnit může [24].

Proč instituce ve světě začaly v uplynulém roce tak masivně podporovat myšlenku open dat?

Důvody:

1. Otevřená data jsou důležitým a nedílným prvkem konceptu transparentní veřejné správy.
2. Odborná veřejnost získává podklady pro svobodnou vědeckou a výzkumnou činnost a je tak schopna daleko efektivněji vyvíjet tlak na racionálnější fungování veřejné správy.
3. Odborná veřejnost může svobodně vytvářet softwarové aplikace, které zpřístupňují data širší veřejnosti ve srozumitelné podobě.
4. Veřejná správa šetří prostředky, protože se může věnovat jen tvorbě strategicky důležitých a zákonem daných informačních systémů.
5. Veřejná správa systematizuje sběr a zveřejňování dat. Lépe se odhalují zdroje duplicitních dat. Veřejná správa získává přehled o tom, kde jsou tvořena jaká data. To v důsledku vede k dalšímu šetření prostředků.

Data jsou otevřená, pokud jsou:

1. **Úplná** - data jsou zveřejněna v maximálním možném rozsahu. Rozsah může být definován právním předpisem, usnesením vlády, příp. poskytovatelem dat. Například seznam všech nemovitostí s číslem popisným nebo evidenčním v obci XY, nebo seznam všech památkově chráněných objektů v obci XY.
2. **Snadno dostupná** - data jsou dostupná na internetu a snadno dohadatelná.
3. **Strojově čitelná** - data jsou ve formátu, který je strukturovaný takovým způsobem, že pomocí programové aplikace z nich lze získat vybrané údaje.
4. **Používající standardy s volně dostupnou specifikací (otevřené standardy)** - data musí být ve formátu, který je volně (bezplatně) dostupný pro libovolné použití nebo jsou do takového formátu převoditelná volně (bezplatně) dostupnou aplikací.
5. **Zpřístupněna za jasně definovaných podmínek užití dat (licence) s minimem omezení** - podmínky musí být jasně a zřetelně definovány a zveřejněny.
6. **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 vyžádat i úhradu za mimořádně náročné pořízení dat, pokud si uživatel zpřístupnění těchto dat vyžádá [25].

2.6 Proces publikace otevřených dat



Obr. 9: Proces publikace open dat [25].

Prvním krokem je **analýza a výběr dat k uveřejnění**. Cílem tohoto kroku je analyzovat dostupná data, popsat jejich strukturu a zvolit data, která jsou vhodná zveřejnit jako otevřená data. Následující krok je věnován **výběru vhodného formátu dat**. Tento krok je zaměřen na výběr vhodného formátu dat z formátů, které jsou standardizované a obecně využívány. Využívání široce používaných formátů přispívá ke snadnějšímu využití zveřejněných dat. Dalším krokem je **návrh způsobu přístupu k datům**, jehož hlavní náplní je rozhodnutí, zda mají být data zpřístupněna v podobě stažitelných souborů nebo pomocí webových služeb. Ve čtvrtém kroku dochází k **exportu dat do navrženého formátu**. V následujícím kroku dochází k **publikaci dat**. V rámci kroku dochází k určení vhodné webové prezentace dat a volbě URL, na které budou data dostupná uživatelům. Posledním krokem je **katalogizace dat**. Zde dochází k tvorbě záznamu o zveřejněných otevřených datech v Datovém katalogu, aby potenciální zájemci mohli data snadno vyhledat [25]. Všechny tyto kroky jsou znázorněny na Obr. 9.

2.7 Open data v ČR

I když je myšlenka open dat nová, fenomén open dat veřejné správy se světem šíří rychle. Proto i Česká republika musí vyvinout vlastní aktivitu, pokud nechce zůstat pozadu jak technologicky, tak společensky. Není ale nutné hned provádět revoluci. Zatím postačí někde začít – zprovoznit katalog pro otevřená data (jehož prototyp již iniciativa OpenData.cz zprovoznila) a začít v něm systematicky zveřejňovat některá data některých pilotních měst či úřadů ČR [24].

Iniciativa <http://OpenData.cz> si klade za cíl:

1. Informovat odbornou i laickou veřejnost o principech a výhodách otevřených dat.
2. Realizovat technickou platformu pro otevřená data v ČR.
3. Systematicky spolupracovat s několika institucemi veřejné správy, analyzovat jejich data a pomáhat je zveřejňovat v otevřené podobě [24].

2.8 Datové katalogy

V dubnu 2012 byl schválen Akční plán České republiky „*Partnerství pro otevřené vládnutí*“. Tento akční plán (dále jen *Akční plán OGP*) obsahuje závazek ČR vybudovat oficiální katalog otevřených dat veřejné správy. V rámci plnění tohoto závazku vznikla Koncepce katalogizace otevřených dat VS ČR3, ve které je popsán navržený koncept Katalogu otevřených dat VS ČR [26].

Aby bylo možné využít potenciálu (otevřených) dat veřejné správy, je třeba, aby tato data byla snadno vyhledatelná. Data veřejné správy často nejsou dostupná z jednoho místa, ale jsou publikována na webových stránkách jednotlivých orgánů veřejné správy. Za účelem usnadnění přístupu k datům veřejné správy vznikají **katalogy dat veřejné správy**.

Datový katalog (katalog dat) lze definovat jako „*soubor katalogizačních záznamů*“, kde tyto katalogizační záznamy obsahují metadata popisující určitá data. Příkladem takovýchto katalogů dat jsou portály data.gov1 (USA) nebo data.gov.uk2 (Velká Británie) [26].

Datové katalogy lze členit z několika hledisek. Z hlediska rozsahu zaměření katalogu lze katalogy členit na **lokální, národní a mezinárodní**. Národní katalogy obsahují údaje o datech z celého státu, oproti tomu lokální datové katalogy obsahují údaje o datech, které se vztahují k určitému území, městu nebo jinému územnímu celku. Lze se setkat i s podrobnějším členěním lokálních datových

katalogů.

Příklady národních datových katalogů.

Stát	Katalog	Oficiální	Nástroj	Počet záznamů ⁹
Belgie	http://data.gov.be	Ano	Proprietární	106
Česká republika	http://cz.ckan.net	Ne	CKAN	171
Francie	http://www.data.gouv.fr/	Ano	Proprietární	353259
Irsko	http://ie.ckan.net/	Ne	CKAN	268
Slovensko	http://data.gov.sk/	Ano	CKAN	45
USA ¹⁰	http://www.data.gov/	Ano	Socrata	367637
Velká Británie	http://data.gov.uk/data	Ano	CKAN	8955

Příklady národních datových katalogů[26].

Datové katalogy lze také členit na **oficiální a neoficiální**. Oficiální datové katalogy jsou vytvářeny veřejnou správou a měly by být důvěryhodným zdrojem informací o datech veřejné správy. Nicméně z iniciativy organizací či jednotlivců vznikají i neoficiální datové katalogy otevřených dat veřejné správy.

Příkladem neoficiálního datového katalogu může být experimentální katalog dat VS ČR (Veřejná správa České republiky), který je dostupný na adrese <http://cz.ckan.net> [26].

Funkce poskytované katalogizačními nástroji

Katalogizační nástroje umožňují v současné době nejen vizualizovat katalogizovaná data v podobě grafů nebo na mapovém podkladě, ale pokud jsou data ve vhodném formátu a je umožněn přístup katalogizačního nástroje k datům, umožňují katalogizační nástroje i procházení tabulkových dat (někdy i s možností filtrovat a řadit obsah sloupců obdobně jako v případě tabulkového kalkulátoru). Funkce katalogizačních nástrojů zobrazuje *Obr. 10*.

Datové modely pro katalogy dat

Datové katalogy jsou uživatelům zpravidla dostupné prostřednictvím internetu jako webové portály. Pro automatizované zpracování jejich obsahu, jako je např. hromadná transformace katalogizovaných dat do jiného formátu, je třeba, aby obsah datového katalogu byl dostupný také v dobře strojově zpracovatelném formátu [26].



Obr. 10: Funkce katalogizačních nástrojů [26].

Kvalita obsahu katalogu otevřených dat

Ačkoli počet dat katalogizovaných na britském portálu data.gov.uk roste, více než čtyři pětiny uživatelů portál opouští, aniž by použili některý z odkazů na data. I když mohou být důvody pro tento stav různé, jako možné příčiny mohou být nepřehledná navigace na portálu a nejednotný způsob kategorizace dat.

Kvalita obsahu katalogu otevřených dat je tak jedním z aspektů, které mohou ovlivnit, zda uživatelé budou tyto katalogy skutečně využívat, protože nekvalitní katalogizační záznamy mohou uživatelům komplikovat vyhledávání a práci s otevřenými daty. Základními hledisky pro posuzování kvality dat je jejich dostupnost, přesnost, úplnost a konzistence. Atributy kvality obsahu jsou vypsány v Obr. 11 [26].

Atribut kvality dat		Kvalita obsahu datového katalogu
Dostupnost	v čase	Katalogy otevřených dat mají zpravidla formu webových portálů. Proto by svým uživatelům měly být k dispozici nepřetržitě.
	v místě	Katalogy otevřených dat jsou zpravidla dostupné prostřednictvím internetu. Měly by tak být dostupné z libovolného místa s připojením k internetu.
	v požadované struktuře	Různě zaměřené datové katalogy mohou vyžadovat různou strukturu metadat v katalogizačním záznamu. Nicméně porovnání datových modelů pro datové katalogy uvedené v tomto článku ukazuje, že lze najít určitou minimální množinu atributů, které se budou vyskytovat napříč datovými katalogy. Při volbě struktury metadat lze vyjít z atributů uvedených v DCMi Metadata Terms a v budoucnu také ze struktury DCAT coby doporučení W3C, případně z dalších standardů, které v této oblasti vzniknou.
	v požadovaném formátu	Obsah katalogu otevřených dat by měl být dostupný ve formátu HTML, který umožní lidem číst jeho obsah s využitím běžného webového prohlížeče. Pro zpracování obsahu datového katalogu pomocí aplikací by jeho obsah měl být dostupný také ve strojově dobře čitelném formátu, např. XML. Do budoucna může být využit formát DCAT založený na RDF.
Přesnost		Metadata v katalogizačních záznamech by měla být správná, tj. měla by přesně odpovídat katalogizovaným datům (skutečnosti). V katalogizačních záznamech by tak např. měl být správně uveden původce/poskytovatel dat, měly by být správně uvedeny podmínky užití dat nebo by měly být uvedeny platné URL datových zdrojů.
Úplnost		Datový katalog by měl obsahovat záznamy o všech otevřených datech, která mají být katalogizována. Zároveň by ale záznamy v datovém katalogu neměly být duplicitní, tj. dva různé záznamy by neměly popisovat stejná data.
Konzistence		Záznamy v katalogu otevřených dat by měly být konzistentní. Data by tak např. měla být jednotným způsobem klasifikována, tj. data ze stejné oblasti by měla být označena stejnými koncepty klasifikačních struktur. Jednotně by také mělo být postupováno v případech, kdy hodnoty určitých atributů metadat nejsou známy, aby uživatelé nebyli rozdílným přístupem zmateni.

Obr. 11: Kvalita obsahu datového katalogu [26].

2.9 Film Open data

Zajímavý film Open knowledge foundation - #opendata

Odkaz na film : <http://opendata.cz/?q=cs/film>

3 LINKED DATA versus OPEN DATA

Musí být všechny LINKED data OPEN daty?

Ne všechna linked data jsou open daty, a ne všechna open data budou linked daty. Data mohou být open, ale nemusí být linked [28].

Linked data vs. open data

“Open data is data that can be freely used, reused and redistributed by anyone – subject only, at most, to the requirement to attribute and sharealike.”

- OpenDefinition.org

Open data

Data can be published and be publicly available under an open licence without linking to other data sources.



Linked data

Data can be linked to URIs from other data sources, using open standards such as RDF without being publicly available under an open licence.

Zdroj: Introduction to linked data [27].

4 BIG DATA

4.1 Co jsou Big data?

Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high velocity capture, discovery and/or analysis [33].

Najít přesnou a hlavně jednotnou definici Big dat je prakticky nemožné. Tento fakt sám o sobě hodně naznačuje i celkovou nejednotnost konceptu jako takového, který je vše, jen ne formalizovaný či jakkoli jinak unifikovaný. Definice v úvodu kapitoly naznačuje, že Big data jsou soubor technologií, které mají za cíl správu a analýzu velkého množství převážně nestrukturovaných dat. Takový popis nicméně úplně přesně nevystihuje jeden důležitý aspekt celého konceptu [34].

Za Big data se tedy považují taková data, která nelze ukládat ani zpracovávat tradičními technologiemi. Tato data jsou natolik kapacitně náročná, že je nelze uložit na jeden disk, a je tedy nutná distribuce na více paměťových médiích. Zároveň se o Big datech často mluví jako o věčných datech, neboť teoreticky není potřeba žádná data mazat, pouze přidávat nová, a zpřesňovat tak výsledky. Naprosto klíčový je tak požadavek rozšiřitelnosti tohoto úložného prostoru za chodu a dle potřeby (teoreticky až do nekonečna). Výpočty nad těmito daty se pak musí provádět distribuovaně. Koncept Big dat mění i samotný postup výpočtů, kdy se již nepřesouvají data na jedno místo kvůli výpočtu, ale výpočet se pak pomocí speciálních algoritmů Map & Reduce distribuuje blíže k datům. Výsledků se pak dosáhne postupným redukováním mezivýsledků [35].

Druhým důležitým benefitem konceptu Big dat je fakt, že celé řešení lze vystavět pouze s využitím komoditního hardwaru bez nutnosti zapojovat drahá proprietární hardwarová řešení či servery.

Příkladem mohou být společnosti jako Google či Facebook. Mají svá datová centra sestavena z relativně levných počítačů zapojených v obrovském počtu paralelně, což jim dodává potřebný výkon. Při výpadku jednoho uzlu jej lze snadno nahradit jiným, který převezme jeho práci [36].

Hlavní rozdíl oproti běžným analytickým přístupům je nicméně v samotném přístupu při pokládání dotazů. Big data tím, že uchovávají a procházejí opravdu všechny dostupné informace v jejich původní podobě bez ohledu na formát, umožňují pokládat prakticky jakékoli otázky. Big data nicméně nemění nic na obecných principech analytické práce, pouze v některých jejích aspektech

podstatně rozšiřují možnosti.

Pojmy jako big data, velký objem dat, rychlost jejich vzniku nebo doba zpracování, jsou v dnešní moderní době využívání elektronických dat, takřka ve všech odvětvích poskytujících produkty a služby, považovány za kritické a prioritní k zamyšlení. Protože žádný malý, či středně velký podnik nebo i velká korporátní společnost by neměly exponenciální nárůst dat přehlížet, ba naopak ho akceptovat a efektivně využít ve svůj prospěch [36].

Data, informace i znalosti jsou nedílnou součástí každého informačního systému. Aktuálním požadavkem na informační systém je centralizace dat, která snižuje nároky na jejich zabezpečení a zálohování. Zatímco se data hromadí, zálohují a centralizují, vyvstává nespočet otázek nad jejich efektivním uložením a zpracováním. Moderní přístupy vyzdvihují *cloud computing* jako účinnou podporu komplexního informačního systému. Lídři v oblasti informačních technologií paralelně vyvíjejí hardware i software pro nalezení vysoce výkonných řešení, mezi které patří např. NoSQL databáze nebo distribuované souborové systémy [36].

Cloud computing

Cloud computing se jeví jako důležitý moderní IT model pro přístup a distribuci výpočetních zdrojů a jeho přínos sumarizují následující body:

- Dovoluje uživateli rozložit svoje výpočetní prostředky (počítače, síť, datová úložiště atd.) bez zbytečného zdržení a komplikací typicky spojenými se získáváním zdrojů.
- Identifikuje různorodé potřeby jednotlivých uživatelů ve společnosti a snáze uspokojuje poptávku po systémových zdrojích.
- Umožňuje společnosti rychle reagovat na měnící se podmínky podnikání – zaměstnanec dané společnosti nepotřebuje pokročilé znalosti v oblasti IT.

Jedním ze způsobu využití je významná open source platforma, jejímž prostřednictvím lze vybudovat vlastní privátní cloud. Na druhé straně stojí veřejné cloudy poskytující aplikace nebo infrastrukturu formou služby. Pojem „velikost“ dat je chápán nejen z hlediska objemu dat měřeného giga-, tera- či petabyty, ale i z hlediska rychlosti jejich tvorby a přenosu a z hlediska různorodosti jejich typů. Jako příklad je často citováno množství údajů o počasí, které získává každý den Národní úřad pro oceán a atmosféru (NOAA) nebo NASA. V souladu s vládními nařízeními a s postupnou digitalizací narůstá objem archivovaných elektronických dokumentů, e-mailových zpráv a dalších záznamů o elektronické komunikaci [37].

4.2 Základní charakteristiky

Ačkoli ohledně přesné definice Big dat panují stále dohady, ustálily se obecně některé charakteristiky, které mají všechna Big data společná a které je odlišují od těch "běžných", zvládatelných dat.

Většina zdrojů uvádí charakteristiky **Volume**, tedy množství dat, **Velocity**, tedy rychlost, jakou data přibývají, a v neposlední řadě **Variety**, tedy různou podobu a formát dat. K těmto třem základním charakteristikám (**3V**) společnost International Business Machines Corporation (IBM) identifikovala ještě jednu – **Veracity**. Charakteristika, která řeší především úplnost a důvěryhodnost zpracovávaných dat. Další charakteristiky mohou postupně přibývat, což dokládá, že i samotný proces poznání Big dat je stále ještě na začátku [36].

1. Volume – Objem

Charakteristika objemu reprezentuje celkové množství dostupných dat. To může být tvořeno historickými záznamy zákazníka, všemi jeho transakcemi, logy jeho aktivit, daty ze sociálních sítí a webu obecně nebo všemi zmíněnými plus dalšími zdroji dohromady.

Typický dosavadní přístup při analytické práci nad daty byl pomocí vzorkovacích algoritmů vybrat určitou reprezentativní podmnožinu dat a na ní testovat určité hypotézy, či dělat statistiky. Z takto získaných výsledků byla často odvozována konkrétní doporučení (jaký produkt si zákazník příště koupí a podobně). Klíčovým bodem v tomto standardním procesu je přitom již samotný výběr vzorku, což je velice složitá operace. Téměř každé vzorkování totiž zavede určitou chybu, či zkreslení, které může ovlivnit (a velice pravděpodobně opravdu ovlivní) výsledek. Rozdíl přístupu Big dat v této oblasti je právě v možnosti zpracovávat všechna data bez ohledu na jejich množství a být si tak jistý, že výsledek je opravdu nejlepší možný a odpovídá skutečné situaci. Jelikož například nebylo možné efektivně využít dosavadní technologie ukládání dat, vznikl čistě pro potřeby zpracování Big dat zcela nový souborový systém **Hadoop**, na jehož základě staví většina současných open source i proprietárních komerčních Big dat řešení [34].

Je třeba zmínit, že charakteristika objemu je vysoce subjektivní a nelze přesně určit, jaké množství dat je již považováno za Big data. Hranice mezi Big daty a normálním množstvím je tenká a neustále se pohybuje. Finální číslo se prakticky pro každou organizaci liší a je dáno i tržním sektorem, ve kterém firma působí, či softwarovými a hardwarovými nástroji, které k jejich analýze využívá. Obecně se ovšem má za to, že Big data jsou někde v rozmezí od několika terabajtů po desítky petabajtů [34].

2. Velocity – Rychlost

Rychlostí je míněna především celková doba od vytvoření nové informace přes její získání, zpracování až k samotnému rozhodnutí. Aspekt rychlosti ke zpracování dat přibyl hlavně v posledních letech a s rozvojem technologií bude nabývat stále větší významu. Současné algoritmy pro automatické obchodování na burze pracují v řádu milisekund a uzavírají obchody za dobu kratší, než člověk stihne kliknout na myš. Milisekunda rozdílu při provedení příkazu v tomto případě znamená zisk nebo ztrátu [34].

3. Variety – Rozmanitost

Klíčovým rozdílem Big dat je především rozmanitost informací, které skrze tuto platformu protékají a kterých se celý koncept obecně týká. Tradiční data, se kterými se při běžném provozu pracuje, mají pevně danou strukturu (obecně se označují pojmem strukturovaná). Typickým příkladem jsou data ukládaná do databázových struktur. Během posledních let nicméně začal převládat obsah nestrukturovaný a především semistrukturovaný. Nestrukturovaný obsah může mít mnoho podob a jen těžko by jej šlo do detailu definovat. Do této kategorie nicméně spadá většina moderních formátů od obrazových či zvukových dat, příspěvků na sociálních sítích, blogů přes informace o geografické poloze, záznamy kliknutí na webech, data dostupná na internetu a další. Všechny tyto formáty mají společný fakt, že nemají pevně daný řád. Často obsahují větší množství textové informace, která přitom nemusí být pouze ve formě srozumitelných vět, ale mohou to být různé logy ze zařízení, záznamy pohybu mobilních telefonů, různé signály, kódy a podobně. Právě logy jsou příkladem informací, které se často označují jako semi-strukturované. Semistrukturovaná data mohou mít v některé své části určitou strukturu (lze je předvídatelně číst), větší částí je ovšem nestrukturovaný text. Úkolem Big dat je vytěžit informace a sjednotit je do podoby vhodné pro další zpracování (do jisté míry v nich jakousi, byť virtuální, strukturu vytvořit) [34].

4. Veracity – Věrohodnost

Společnost IBM definovala ještě jednu charakteristiku, která vznikla především na základě zkušeností z praxe. V tradičních databázových systémech bylo (a stále je) věnováno hodně pozornosti přípravě a předzpracování dat ještě, než vstoupí do samotného systému, a následně pak analýz. Připustíme-li v kontextu Big dat myšlenku zapojení dat z různých sociálních sítí, webových článků a jiných podobných zdrojů, je logické, že konzistence a důvěryhodnost těchto dat může z pohledu analytiků poklesnout. Ve skutečnosti dokonce dvě z výše zmíněných charakteristik, kterými jsou rychlost a rozmanitost, doslova pracují proti věrohodnosti. Pokud totiž v reálném čase

zpracováváme opravdu velké množství dat z různých zdrojů, není na jejich důkladné čištění nebo filtrování příliš času. Dopady záleží především na tom, jak daná organizace takto získaná data využívá a jaká rozhodnutí z nich chce vyvodit. Nemusí to nutně znamenat nevýhodu, je pouze potřeba si tuto skutečnost uvědomit [34].

4.3 Způsoby zpracování velkého objemu dat

1. Apache Hadoop

Apache Hadoop je opensource framework pro ukládání a zpracování dat bez ohledu na jejich formát. Počátky technologie sahají až k firmě Google, která si pro potřeby svého vyhledávače vytvořila vlastní souborový systém Google File System (GFS). Důležitým aspektem zde přitom byl princip zpracování dat pomocí modelu Map&Reduce algoritmů, které umožnily vysokorychlostní zpracování paralelních dat. Koncepty souborového systému Google převzali dva vývojáři společnosti Yahoo a vytvořili nástroj Hadoop, který pak zadarmo dali k dispozici veřejnosti jako open source [34].

Hadoop se sestává z dvou klíčových komponent:

- **Hadoop Distributed File System (HDFS)**
- **Map&Reduce**

HDFS

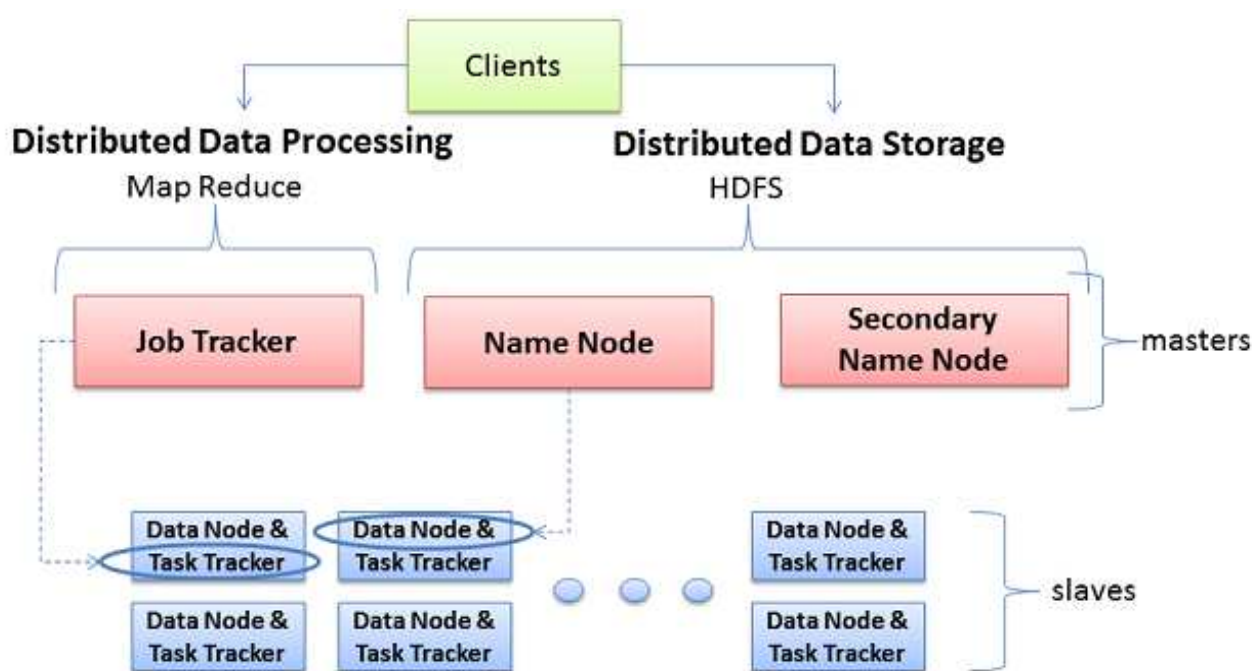
HDFS je distribuovaný souborový systém navržený k držení velmi velkého objemu dat, který svojí vysokou propustností poskytuje přístup klientovi ke všem uloženým informacím. Soubory jsou uloženy redundantním způsobem, aby byla zvýšena odolnost proti chybám a umožněna práce paralelně spuštěným aplikacím. HDFS je založen na konstrukci Google file system (GFS) a je navržen tak, aby byl odolný co nejvíce problémům [36].

Vyznačuje se zejména těmito vlastnostmi:

- poskytnutí rychlého, stálého a škálovatelného přístupu k velkému množství informací,
- šíření dat v rámci velkého počtu strojů (uzlů) začleněných do jednotlivých clusterů,
- cluster odolný kompletnímu selhání několika strojů, aniž by ztratil informaci nebo výrazně zpomalil výkon.

Map&Reduce

Velké úlohy jsou rozděleny na menší úkoly (text rozdělen do jednotlivých vět), které jsou následně paralelně zpracovány jednotlivými procesory. Výsledky paralelního počtu jsou následně seřazeny a redukovány do výstupní množiny. Hadoop přijímá požadavky na jeden centrální (master) uzel a následně je distribuuje jednotlivým uzlům v clusteru funkcí MAP, která rozdělí vstupní požadavek na menší podproblémy pro jednotlivé instance. Uzel tuto funkci vyhodnotí a vrátí výsledek master uzlu. Ten pak má na starosti agregaci takto získaných výsledků funkcí Reduce a navrácení výsledného subsetu dat uživateli. Funkce Map&Reduce se obvykle programuje pomocí jazyka Java nebo C++. K ovládání Hadoop souborového systému tímto způsobem je tedy potřeba poměrně vysoká technická znalost [36].



Obr. 13: Schéma Hadoop systému.

Hlavní výhodou Hadoopu je, že umožňuje analyzovat úplné datové soubory údajů, včetně nestrukturovaných a částečně strukturovaných dat, a to z hlediska nákladů i času efektivním způsobem. Mezi nevýhody Hadoopu patří částečná nezralost a hektický vývoj. Kromě toho, zavádění a řízení Hadoop clusterů a provádění pokročilé analýzy na velké objemy dat vyžaduje značné odborné znalosti. Pro firmy je takový model vesměs nepřijatelný, a proto v rámci ekosystému vznikla řada firem, které staví komerční řešení na bázi Hadoopu tak, aby se nasazení a správa technologie stala praktickou realitou tradičního enterprise odvětví [36].

2. Apache Cassandra

Je masivně škálovatelná NoSQL databáze navržená tak, aby zvládla práci s velkým objemem real-time dat na několika datových centrech s žádným kritickým bodem pro selhání celého systému a poskytovala podnikům vysoký databázový výkon a neustálou dostupnost. Hlavním komerčním distributorem software, technické podpory a školení je společnost DataStax. Cassandra na rozdíl od Hadoop nepracuje jako key – value (klíč – hodnota) úložiště, nýbrž jako column – family (sloupec – rodina) úložiště, kde se ukládaný objekt skládá z trojice: název, klíč, časový otisk. V praxi to znamená upřednostnění využití projektu Cassandra na úkor Hadoop při zpracování dat z naměřených časově podložených údajů nebo bezpečnostních logů, kde čas hraje významnou roli [36].

4.4 Příklad Big Dat

China Telecom

Zajímavým příkladem z pohledu objemu zpracovávaných dat je implementace řešení pro největšího mobilního operátora na světě, China Telecom. Tato firma má více než 600 milionů klientů rozdělených mezi 31 samostatných poboček. Data generovaná klienty i samotným provozem sítě za rok 2012 dosáhla kolosálních rozměrů. Podpůrné systémy pro služby vygenerovaly 8000 TB, data primárně nasbíraná a určená pro pozdější business analýzu 7000 TB, log detailních informací o hovorech všech zákazníků 16 TB a záznamy pohybu všech zařízení podle jejich signálu generují pravidelně více než 1 TB denně. Tradiční centralizovaná řešení sestavená z drahých serverů byla nahrazena levnější cloud computing variantou založenou na třech datových centrech postavených na kombinaci lacinějšího hardwaru a tradičních Big datech technologiích, jako Hadoop a Map&Reduce pro distribuci a kolekci výpočtů nad daty. Všechna tato data jsou nyní přístupná k analýze, reportům a jakémukoli dalšímu využití. Značně se také zkrátila doba předzpracování dat, neboť je nyní možné ukládat je přesně v tom formátu, v jakém vznikla, což by předtím nebylo možné [34].

4.5 Big data a GIS

Vzrůstající nároky na datový objem, druh i rychlost, stejně jako na jejich kvalitu, kladou různé otázky pro výzkum ohledně ukládání, analyzování a zpracování dat. Kromě technologických otázek je nutné zaměřit se na management, optimalizaci podnikových procesů a zjišťování ekonomické hodnoty podnikových modelů operujících s Big daty. Big data odpovídají enormnímu objemu, druhu a dostupnosti dat, přičemž efektivní přístup se může změnit v konkurenční výhodu podniku.

Mnoho podniků akumuluje astronomický rozsah dat, který již dávno překročil alarmující meze. Objem veškerých digitálních dat vzrostl v roce 2012 na 2,7 miliard terabytů, meziročně nárůst o 48%, a podle prognóz dosáhne hranice 8 miliard terabytů v roce 2015. V současné době lze IT svět rozdělit podle přístupu společností k datům na podniky, které data ukládají a skladují, jiné data dolují pro jejich pochopení a další data agregují pro snazší interpretaci. Shromažďovat, standardizovat nebo očišťovat data od nesrovnalostí v takto velkém rozsahu předkládá podnikům další překážky, jejichž odstranění se stává prioritou [36].

Významnou roli v mnoha oblastech moderního světa hrají geografické informační systémy (GIS), kde by ztráta kontroly nad daty omezila nebo zpomalila vývoj, implementaci i využití GIS, a to by s sebou např. v případě krizového řízení neslo fatální důsledky. Zhruba 80% všech informací obsahuje prostorovou komponentu. Spousta důležitých funkcí je tak závislých na prostorových datech. Proto vznikají nové integrační služby nebo služby spojené se sdílením prostorových informací. Existuje mnoho faktorů, jako drahé náklady na software, nákladná údržba, malý hardwarový výkon nebo nízká systémová bezpečnost, které brzdí vývoj služeb pracujících s prostorovou informací. Příkladem využití GIS při práci s prostorovými daty mohou být funkce spojené s environmentálním plánováním, funkce krizového managementu nebo funkce spojené s demografií [36].

Objemná prostorová data

Na objemnost prostorových dat lze nahlížet ze dvou úhlů pohledu. Jednak je způsobena množstvím údajů v atributových tabulkách, kde se ke každému bodovému záznamu objevují, kromě jednoznačného identifikátoru, jeho souřadnice, k liniovým prvkům lze přidat nebo vypočítat délku a k polygonům rozlohu. Každý geoprvek může obsahovat nepřeberné množství atributů pro různé účely zpracování a v reálném případě několika set tisíc záznamů tvořících mapový list se velikost prostorových dat pohybuje ve stovkách MB až GB. Druhý úhel pohledu tvoří kvalita a rozměr pořizovaných snímků zemského povrchu. Zatímco velikost snímku do rodinného alba se pohybuje okolo 5–10 MB, originální velikost leteckého snímku ve formátu tiff. může být až desetinásobná [36].

Trendy pro zpracování velkých objemů prostorových dat

Jedním ze zástupců velkého objemu prostorových dat je shromažďování záznamů o pohybu lidí. Např. středně velká datová sada obsahující stovky uzlů v případě migrace obyvatel v České republice, USA nebo jiných zemích zahrnuje tisíce cest a směrů pohybu, jež chtějí být zkoumány. Vznikají ovšem daleko větší soubory v podobě přehledů a simulačních modelů nad epidemií

zasaženými oblastmi. Již k tak velkému objemu dat je třeba přičíst i množství nejrozličnějších proměnných v podobě různých skupin věku obyvatelstva, pohlaví, atd. Pro vizualizaci prostorových interakcí se využívá **flow mapping**. Flow mapy pracují nad datovými sadami a využívají se k zobrazení migračních pohybů nebo k zaznamenávání letecké dopravy v časových intervalech [36].

S popularitou GIS v mnoha organizacích roste i množství prostorových dat ukládaných v databázích nebo jiných informačních úložištích. V důsledku toho se objevují nové výzvy pro vytváření inovativních metod pro extrahování informace z velkého objemu dat. Data mining nabízí několik metod pro dolování znalostí z rozsáhlých databází nebo datových skladů. Pro rozsáhlé analýzy dat se využívají datové kostky, ve kterých jsou data uspořádána podle důležitosti jejich atributů pro konkrétní analýzu. Kostka je tvořena naměřenými hodnotami, které jsou kategorizovány do dimenzí. Dimenze jsou uspořádány do hierarchií s agregovanými daty podle zvolené rozlišovací úrovně. Ačkoliv míry ani dimenze nejsou prostorovými údaji, lze zavést nový datový model spatial data cube (datová kostka pro práci s prostorovými daty), ve kterém míry i dimenze podporují jak neprostorová, tak i prostorová data.

Dotaz na spatial data cube by mohl vypadat např. takto: „Ukaž evropské státy s podobnou teplotou v jarních měsících.“ Státy představují prostorové míry, protože podle jejich prostorového umístění bude zjištěna teplota [36].

Technologie **grid computing** nabízí další způsob, jak zvládnout současnou poptávku po úložných a výpočetních kapacitách. Grid je druh infrastruktury charakterizované sdílením zdrojů. Grid computing umožňuje sdílet kapacity a řešit problémy spoluprací v rámci distribuovaného multiorganizačního prostředí. Při práci s prostorovými daty lze grid computing uplatnit jak na víceúrovňová data v rastrové reprezentaci, tak na geometricky přesnější vektory. Všechna data jsou tradičně ukládána v prostorových databázích, vždy podle současných požadavků. Při práci s prostorovými daty nabízí grid computing využití široké škály služeb, prohlížením a dotazováním nad ekonomickými statistickými daty počínaje a placenými službami jako upravování dat využitím superpočítače v rámci jednoho či více uzlů v grid architektuře konče [36].

4.6 Uživatelé Hadoop

Seznam uživatelů Hadoop, odkaz: <http://wiki.apache.org/hadoop/PoweredBy>

4.7 Konference Big data

Konference BIG DATA, odkaz: <http://www.ieeebigdata.org/2013/research.html>

5. REFERENCE

Linked data

- [1] Bizer Ch., Heath T., Berners-Lee T., *Linked data A story so far*, Freie Universität Berlin, Germany, Talis Information Ltd, United Kingdom, Massachusetts Institute of Technology, USA, 2009 [online], [cit. 2013-11-21] Dostupný z WWW: <<http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>>.
- [2] Berners-Lee T., *Linked Data – Design Issues*, W3C [online]. [cit. 2013-11-28] Dostupný z WWW: <<http://www.w3.org/DesignIssues/LinkedData.html>>.
- [3] Mynarz J., Zemánek J. *Úvod k linked data*. Knihovna plus [online]. [cit. 2013-11-28]. Dostupný z WWW: <<http://knihovna.nkp.cz/knihovnaplus101/myna.htm>>. ISSN 1801-5948.
- [4] Pánek O., *Využití principu Linked Data pro účely účetních výkazů měst a obcí*. Bakalářská práce [online]. [cit. 2013-11-28]. Dostupný z WWW: <https://dip.felk.cvut.cz/browse/pdfcache/panekon1_2012bach.pdf>.
- [5] Bizer Ch., Heath T., *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool Publishers 2011. s. 1-136. ISBN 978-160-845-4310.
- [6] W3C -Semantic web, *What is Linked Data?* W3C [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.w3.org/standards/semanticweb/data>>.
- [7] Satrapa P., *HTML v příkladech*, Softwarové noviny číslo 4/97 [online]. [cit. 2013-11-22] Dostupný z WWW: <<http://www.nti.tul.cz/~satrapa/docs/wwwprikl/html9.html>>.
- [8] Daconta, M. C., Obrst, L. J., Smith, K. T.: *What Is the Resource Description Framework?* [online]. [cit. 2013-11-22] Dostupný z WWW: <<http://www.devx.com/semantic/Article/34816>>.
- [9] [GlobalSemantic] Bartoš, P.: *Sémantický web (GlobalSemantic.net)* [online]. [cit. 2013-11-25] Dostupný z WWW: <<http://sites.google.com/a/globalsemantic.net/gsn/swp>>.
- [10] W3C -Semantic web, *Cool URIs for the Semantic Web* [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.w3.org/TR/cooluris/>>.
- [11] GeoNames [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.geonames.org/>>.
- [12] Hellmann S., Auer S., Lehmann J., *Linkedgeodata - adding a spatial dimension to the web of data*. In Proceedings of the International Semantic Web Conference, 2009. [online]. [cit. 2013-11-29] Dostupný z WWW: <http://link.springer.com/chapter/10.1007%2F978-3-642-04930-9_46>.
- [13] BBC [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.bbc.co.uk/music>>.
- [14] Data.gov.uk, *What is linked data?* [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://data.gov.uk/linked-data>>.
- [15] Hartig, O., *Provenance Information in the Web of Data*. Proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009).
- [16] Cyganiak, R., Delbru, R., Stenzhorn, H., Tummarello, G., Decker, S., *Semantic sitemaps: Efficient and flexible access to datasets on the semantic web*. In Proceedings of the 5th European Semantic Web Conference, 2008. s. 690-704. ISBN 3-540-68233-3 978-3-540-68233-2.

[17] Alexander K., Cyganiak R., Hausenblas M., Zhao J., *Describing linked datasets*. In Proceedings of the WWW 2009 Workshop on Linked Data on the Web, 2009 Dostupný z WWW: <<http://core.kmi.open.ac.uk/download/pdf/10851951.pdf>>.

Open data

[18] Otevřená data [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.otevrenadata.cz/>>.

[19] Auer, S. R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. *"DBpedia: A Nucleus for a Web of Open Data"*. The Semantic Web. Lecture Notes in Computer Science 4825. p. 722, 2007. ISBN 978-3-540-76297-3.

[20] Deirdre L. *Open Data Overview*. DERI, NUI Galway.

[21] Open Knowledge Foundation: *Open Government Data*. [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://opengovernmentdata.org>>.

[22] Open Knowledge Foundation: *Open Data Handbook Documentation* [online]. [cit. 2013-11-17] Dostupný z WWW: <<http://opendatahandbook.org/en/index.html>>.

[23] ESRI, *Open Data - Share and map your data with ArcGIS® technology*. [online]. [cit. 2013-11-25] Dostupný z WWW: <<http://www.esri.com/library/brochures/pdfs/open-data-gov20.pdf>>.

[24] Open data, *Veřejná správa jako zdroj dat* [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://opendata.cz/cs/node/29>>.

[25] Chlapek D., *Metodika publikace otevřených dat veřejné správy ČR*. Praha, listopad 2012 [online]. [cit. 2013-11-30] Dostupný z WWW: <http://www.korupce.cz/assets/partnerstvi-pro-otevrene-vladnuti/otevrena-data/Metodika_Publ_OpenData_verze_1_0.pdf>.

[26] Kučera J., *Katalogizace otevřených dat veřejné správy*. Vědecká konference doktorandů - únor 2013. [online]. [cit. 2013-11-30] Dostupný z WWW: <http://www.ondrejsimpach2.ic.cz/publikace/konference_mezinarodni/FIS_VSE_DEN_DOKTOR_ANDU2013/prispevky/INF_KUCERA_DD_FIS_2013.pdf>.

[27] Introduction to linked data [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.slideshare.net/OpenDataSupport/introduction-to-linked-data-23402165>>.

[28] Linked data [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://linkeddata.org/>>.

[29] Open definition [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://opendefinition.org/>>.

[30] Šošolík I., *Textové formáty* [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.zsjablunka.cz/html/vyuka/informat/text/formaty.pdf>>.

[31] Úvod do JSON [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.json.org/json-cz.html>>.

[32] Common Format and MIME Type for Comma-Separated Values (CSV) Files [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://tools.ietf.org/html/rfc4180>>.

Big data

- [33] Gantz D. R, J., *Extracting value from chaos*. Červenec 2011 [online]. [cit. 2013-11-30] Dostupný z WWW: <<http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>>.
- [34] Tříška M., *Customer Intelligence v kontextu Big Data*. Diplomová práce, 2013 [online]. [cit. 2013-11-26] Dostupný z WWW: <<http://www.diplomovaprace.cz/61/diplomka.pdf>>.
- [35] Fisher D., DeLine R., Czerwinski M., Drucker S., *Interactions with big data analytics*. *Interactions*. Str. 50–59, květen 2012 [online]. [cit. 2013-11-26] Dostupný z WWW: <http://research.microsoft.com/pubs/163593/interactions_big_data.pdf>.
- [36] Špidlen J., *Zpracování velkých objemů prostorových dat*. Diplomová práce, 2013 [online]. [cit. 2013-11-26] Dostupný z WWW: <https://dspace.upce.cz/bitstream/10195/52546/3/SpidlenJ_ZpracovaniVelkych_JK_2013.pdf>.
- [37] Dolák O., *Big data, Nové způsoby zpracování a analýzy velkých objemů dat*. [online]. [cit. 2013-11-26] Dostupný z WWW: <<http://www.systemonline.cz/clanky/big-data.htm>>.