# Budoucnost vyhledávání: Mezi soukromím, technologií a legislativou

**Bc. Michal Černý**
Přírodovědecká fakulta MU Brno
cernymichal@mail.muni.cz

INFORUM 2012: 18. konference o profesionálních informačních zdrojích

Praha, 22. - 24. 5. 2012

---

**Abstrakt:** Efektivní vyhledávání dat a informací je jednou z nejdůležitějších činností, bez kterých se informační společnost nemůže obejít. V zásadě všichni vnímají jako velký problém dolování dat, neboť počítače nemají schopnost textu porozumět a pochopit jej. Proto se začal stále intenzivněji budovat koncept sémantického webu, který se snaží užít XML formáty k tomu, aby bylo možné vytvořit datovou strukturu, která by byla pro počítače co možná nejpřehlednější a umožnila jim odvozování nových znalostí nebo práci s přirozeným jazykem.

Do určité míry se tak otázky vyhledávání informací staly problémem čistě technologickým. Je ale třeba si uvědomit, že celá problematika je podstatně složitější. Předně je třeba vzít v potaz legislativní prostředí, které je v oblasti práce s informacemi mimořádně složité a nepřehledné. Vyhledávání informací je otázkou z hlediska práva velice komplikovanou také proto, že zasahuje do oblasti mezinárodního práva, ale také do legislativních norem všech stát, které v procesu vyhledávání informací libovolným uživatelem vystupují. Současně je třeba vidět ještě jedno omezení, které je na úrovni každého jednotlivce, jakožto lidské osoby. V západní civilizaci má většina lidí za to, že soukromí je hodnota, která má mimořádnou cenu a je třeba ji bránit.

# Úvod

Současná společnost bývá často označována jako znalostní či informační. Jsou to právě informace, které se staly primárním ekonomickým statkem všech vyspělých ekonomik a které se snažíme také patřičným způsobem legislativně chránit. Farmaceutické formy či softwarové společnosti nemají svůj zisk založený na množství či kvalitě strojů, ale na tom, do jaké míry jsou schopny pracovat s daty, informacemi a znalostmi ke svému prospěchu. Ostatně letmý pohled do žebříčků, které zachycují nejbohatší osobnosti světa či nejcennější značky na trhu je relativně výstižný a každý rok se stále více objevují v popředí společnosti, které „obchodují“ s informacemi.

Současně s tím můžeme sledovat poměrně složitou a zajímavou diskusi nad tím, do jaké míry mají být informace dostupné všem a v jakém rozsahu. Není tomu tak dávno, co jsme mohli sledovat kauzu okolo Wikileaks. Tato organizace se rozhodla důsledně dodržovat princip transparentnosti informací za každou cenu – bez ohledu na to, zda ohrozí bezpečnost osob či států, ekonomicky či lidsky někoho poškodí nebo bude mít její chování libovolné jiné nedobré důsledky. Samotný princip dostupnosti dat se stal kategorickým imperativem, podivně vytrženým z kontextu okolního světa.

V akademickém světě se můžeme setkat s řadou konceptů otevřeného sdílení vědeckých poznatků a akademické práce, které začínají veřejně dostupnými diplomovými pracemi, pokračují přes aktivity jako je OpenAccess a končí projekty jako je xArch, které umožňují přístup k moderním vědeckým poznatkům z řady oborů široké veřejnosti.

Stojíme ale před otázkou mnohem širší než jen prosté zpřístupnění dat. Těch je k dispozici tolik, že jejich snadná orientace v nich není vůbec snadná. Proto mají stále větší význam technologie, které je umí hromadným způsobem strojově zpracovávat a analyzovat. V tomto kontextu se často hovoří o dolování dat či zpracování přirozeného jazyka. O technologiích, které by nám umožnily lépe poznat co je obsahem daného dokumentu a lépe s ním pracovaly.

Existuje přitom netriviální množství dat, která se v současnosti rozumným způsobem zpracovávat nedají a přesto jsou jejich prostřednictvím (z různých důvodů) publikovány zajímavé informace. Zde jde především o velice kontroverzní technologii Flash, která je ale (zdá se) na ústupu.

Zpracování informací s sebou pak může přinášet řadu zajímavých a závažných problémů. Předně jsou zde informace chráněné legislativně, kde je jejich zpracování a šíření silně omezováno. V této oblasti neslavně prosluly legislativní normy jako jsou ACTA, SOPA, PIPA a řada dalších, které se pokoušely dostupnost určité množiny dat na internetu omezit. Praktické důsledky jejich implementace by ale měly závažný vliv na svobodu, demokracii i hospodářskou soutěž. Vedou nás proto spíše k hlubšímu promýšlení konceptu autorského práva, které bude muset doznat zásadních

změn. V této oblasti jsme ale stále spíše na počátku.

Zcela samostatnou kapitolou jsou otázky ochrany soukromí. Je třeba říci, že soukromí je jednou z nejvýznamnějších hodnot, kterou si západní civilizace vydobyla a jistě není náhodou, že o jeho omezování usilovaly v minulosti výhradně totalitní režimy či fanatikové. S tím, jak se do popředí zájmu dostává také počítačové zpracování emocí, dostává ochrana soukromí zcela nový, širší a daleko závažnější rozměr.

V tomto příspěvku se pokusíme o postupné rozklíčování tří základních pilířů (byť jen ve velice hrubých rysech), které jsou s problematikou vyhledávání dat a informací spojené – tedy s technickými problémy a omezeními, legislativními normami a ochranou soukromí, na které se velice často zapomíná.

# Technologická omezení - sémantická data v nedohlednu

Vede-li se diskuse o vyhledávání informací, zřejmě největší naděje jsou vkládány do konceptu sémantického webu či desktopu.[1] Oba jsou založeny v zásadě na stejné myšlence. Vytvořit databází znalostí, ve které budou udržovány nejen jednotlivé entity, ale také vztahy mezi nimi. Princip odvozování nových informací by byl řešen (alespoň do určité míry) prostřednictvím logického odvozování, v současné době však naráží na velká omezení. Například stále není jasné, zda se podaří vytvořit algoritmus, který by byl schopen dokázat (byť jednoduchou) matematickou větu.

V budoucnu je možné předpokládat masivní rozvoj umělé inteligence, která by uměla provádět stále lepší rezoluci (v rámci logického programování). Technologicky je možné v pozadí již zde vidět první problém. Velké databáze není možné tvořit pomocí SQL databázových systémů, které jsou spolehlivé, komplexní, ale nepříliš rychlé. Proto jsou často doplňovány (či nahrazovány) NoSQL databázemi, které umožňují pracovat s daty se složitou vnitřní strukturou (ty v SQL nejde rozumným způsobem nijak zpracovat), užívá se model key-value (jednoduché, rychlé, ale stereotypní), síťové databáze (pracují s entitami a jejich vztahy) a řada dalších.

Tyto databáze sice fungují a běžně se užívají (Red Hat, LinkedIn, Twitter, Facebook a další), ale ukazuje se, že daň za strukturovaně uložená data je malá rychlost, za rychlé databáze, pak možnost vyhledávat jen podle jednoho klíče, tedy malá flexibilita. V zásadě je možné říci, že na úrovni databází znalostí není zatím žádné komplexně funkční řešení, které by bylo použitelné pro sémantický web. Za rychlost platíme ztrátou komplexnosti. Proto se NoSQL a SQL databáze kombinují, což ale není dlouhodobě zřejmě udržitelný koncept.

Druhým problémem je, že se autorům nechce sémantické informace k dokumentům připojovat. Je

to pomalé, finančně neefektivní a proto zatím zbytečné. Známý koncept RDF, který měl být krokem k novému webu skončil fatálním fiaskem pro svoji složitost, nikoli pro praktické technické problémy. O konceptu sémantického webu se mluví již hodně dlouho a zatím je možné vidět jen několik málo dílčích úspěchů, které spíše připomínají jeden z pilířů sémantického webu - prostupnost dat skrze mikroformáty (například kalendáře ve formátu iCal).

Budoucnost je také možné vidět v HTLM 5, do kterého jsou vkládány velké naděje. To nabízí především dobré oddělení obsahu od vzhledu. Díky atributům bude možné snadno odlišit článek, rozhovor a třeba navigační menu či patičku webu. Jde o velice dobrý krok směrem ke strojovému zpracování dat, ale také pro přístupnost webu nevidomým osobám či slabozrakým.

HTML 5 tak představuje jednoznačně pozitivní pokrok v tom, jak je možné s webem v budoucnu pracovat - bez Flash, ze kterého nikdo nic strojově nepřečte, bez designu plného tabulek, ve kterých se nemá šanci stroj vyznat a případně bez dynamicky se měnících skriptů, jenž lze systematicky analyzovat také jen velmi obtížně. Nová technologie umožňuje oddělit vzhled od obsahu (plně) a u konkrétního obsahu navíc velice dobře specifikovat, co je zač.

Jestliže se dnes intenzivně pracuje na prohledávání a analýze multimediálních dat (tagy video, audio, canvas), tak HTML 5 nabízí možnost jednoduchého spojení textu s těmito informacemi do logického celku velice jednoduše a elegantně. Připojuje se tedy alespoň základní struktura k dokumentu, což je pro strojové zpracování mimořádně důležité a užitečné.

Zajímavé bude také sledovat vývoj v teoretické informatice, kde by bylo možné pro účely vyhledávání dat a jejich analýzu používat biologické algoritmy či neuronové sítě. Jde o mimořádně zajímavou oblast, které by se měla také u informačních specialistů věnovat podstatně větší pozornost.

# Legislativní omezení

Zatímco se všude hovoří o Open Access a otevřeném přístupu k informacím, skutečnost je o poznání složitější. Osobně si myslím, že je třeba rozlišit dvě základní skutečnosti. Především jsou to země, které na základě svého politického zřízení (nedemokratického) zabraňují občanům k přístupu k informacím zcela obecně a systematicky. Může to být na základě literární cenzury, technických prostředků (Čínský firewall) nebo blokováním vybraných služeb internetu, které vyhledávání informací umožňují (Čína, Írán, Sýrie a řada dalších). Pro obyvatele těchto států je problémem dostat se k jakýmkoli informacím a legislativní omezení představuje vůbec největší překážku.

Druhou skupinou zemí jsou demokratické společnosti, které z různých důvodů neumožňují přístup ke všem informacím - ať již z důvodu ochrany autorských práv, bezpečnosti či jiných. Na ty se nyní

pokusíme podívat alespoň trochu podrobněji a to především v kontextu zkratek SAPA, PIPA či ACTA, které jsou dnes všeobecně známé a hojně diskutované. Ač je dnes téměř jisté, že nebude platit žádná z nich, má přesto význam se jim alespoň krátce věnovat, neboť naznačují směr, jakým se bude celá problematika do budoucna vyvíjet. Ostatně další normy jako je Euro-ACTA se aktuálně diskutují.

SOPA a PIPA [2] [3] [4] jsou relativně přesné právní dokumenty, jejichž platnost byla omezena na území Spojených států. V zásadě je možné vysledovat několik společných motivů. Předně je to převedení povinnosti aktivně vyhledávat obsah, který porušuje autorská práva na provozovatele webu. V současnosti platí DMCA, která říká, že provozovatel je povinen po upozornění obsah odstranit. Nyní by byl povinen jej sám aktivně vyhledávat a zabraňovat tomu, aby se na webu objevil. Weby, které tak nebudou činit mohou být zablokovány.

To je první velice sporný bod, neboť umožňuje zablokování webu na základě útoku (třeba na diskusní fórum), který by byl následně řešen soudní cestou. Samotné blokování webů mimo území USA by mělo být řešeno na úrovni DNS, což je technicky dost obtížně možné. Ač by bylo zřejmě možné se na weby dostat, významně by to zpomalilo celý internet a možnosti jakéhokoli hromadného zpracování. [4]

Druhým významným bodem byla samotná představa filtrování obsahu. Podle norem by měla ochrana autorských práv (respektive jejich vykonavatelů), přednost před ochranou soukromí. Osobně si myslím, že zde je jeden z největších problémů současných diskusí. Zda je možné materiální hodnotu nadřadit soukromí a svobodě člověka. Můj soukromý názor je takový, že nikoli.

Posledním důležitým bodem, který je důležitý pro zpracování informací (a který byl obsažen také v ACTA) je posílení role DRM (digitální ochrana proti kopírování). V zásadě by nemělo být možné DRM obcházet, popisovat způsoby jejího obejití ani nabízet software, který toto umí. Pokud budou trhu stále více dominovat knihy s DRM ochranou, lze si jen dost obtížně představit jejich digitální zpracování, pokud nebude možný jejich převod do jiného formátu. Celý proces se neúměrně prodlouží, prodraží a jeho efektivita klesne. [2] [3]

Objektivně hovořit ohledně ACTA [5] je v zásadě nemožné, neboť jde o amorfní dokument, který umožňoval řadu konkrétních implementací na úrovni států, čímž se stal naprosto nejasným. Tím, že jej Polsko zamítlo ztratila tato dohoda mezinárodního charakteru na významu.

Pokud půjde o budoucnost vyhledávání a zpracovávání informací, pak je možné říci, že právě legislativní normy budou jedním z největších limitujících faktorů. Ať již ochranou autorských práv, jejich vynucováním, nebo také řadou dalších úprav. Ty umožňují budování profilů na sociálních sítích, které nelze přenést nikam jinam nebo je hlouběji analyzovat.

# Ochrna soukromí

Jak jsme již naznačili, otázka vyhledávání dat se relativně nenápadně, ale o to významněji začíná střetávat s konceptem soukromí [7] a to hned v několika oblastech, které si v této části – opět jen schématicky, načrtneme.

V prvé řadě jde o fenomén digitální identity a digitálních stop. Stále více se ukazuje, že pobyt na internetu není bez rizika. Téměř každá lidská aktivita – vyhledávání, napsání komentáře nebo vytvoření profilu na nějaké sociální síti s sebou přináší řadu datových fragmentů, které je možné sestavit do mozaiky, která bude o člověku vypovídat relativně mnoho. Těmto nechtěně či neplánovaně zanechaným informacím se říká digitální stopy a mohou obsahovat řadu důležitých informací o člověku. V současné době se silně diskutuje, do jaké míry budou takto získané informace analyzovány personalisty či dalšími odborníky, kteří budou moci na základě nich efektivněji pracovat s lidmi.

Ostatně tento trend prosakuje již do běžné lidské činnosti. Pokud narazí na internetu na někoho zajímavého, je zcela přirozené, že se pomocí vyhledávacího stroje pokusí o něm zjistit co možná nejvíce informací. Přirozená lidská zvědavost je tímto způsobem uspokojována a současně nabourává představu mnohých o soukromí, které je ale nezbytným předpokladem svobodné společnosti.

Zajímavé otázky se objevují také v kontextu rozvoje počítačového zpracování emocí a senzorických sítí. Zatímco klasické informace, tak jak byly informačními specialisty vyhledávány v minulosti měly převážně lidské autory, situace se zásadním způsobem mění. Moderní automobily jsou schopné průběžně měřit řidiče a posoudit, zda není příliš unavený na pokračování v jízdě. Pokud je, vůz samotný jej donutí zastavit. Google si nedávno patentoval systém reklamy, který bude založený na studiu šumu v hovorech nebo na poloze a datech ze senzorických sítí.

Zdá se, že budoucnost je nevyhnutelně taková, že počítače budou provádět měření emocí a přesně řídit, kdy mají lidé pracovat, kdy zajít na přestávku nebo si zaběhat. Zaměstnavatel tak může mít mnohem lepší přehled o emocionálním stavu svého zaměstnance než jeho manželka či rodiče. Otevírá se velká otázka, zda a jak s takto získanými daty pracovat. Jde o mimořádně užitečné informace pro psychology, sociology či manažery. Ale také o intenzivní narušení intimního prostoru a soukromí. Tato data se přitom během několika málo let stanou běžnou součástí informačního étosu.

Na tomto místě si dovolíme ještě dva protichůdné příklady, jak se k ochraně soukromí přistupuje.

Na jedné straně je zde budována anonymní P2P síť TOR, která využívá konceptu Onion routing s asymetrickým šifrováním proto, aby zajistila anonymní pohyb na internetu. [8]

Na druhé straně můžeme sledovat skutečnost, že každá barevná laserová tiskárna vytváří na obrazu okem neviditelné matice, které umožňují přesně identifikovat zdroj, ze kterého tiskovina pochází. Podobnost s tím, jak se používali psací stroje u nás či v NDR není rozhodně jen čistě náhodná. To vše se děje za záminkou, že jde o boj proti padělání peněz.

# Závěr

Budoucnost vyhledávání informací není samozřejmě černá. Rozvíjejí se vyhledávače zkoušející zpracovávat přirozený jazyk, pracující s obrázky, videem či zvukem. [6] Existují ale limity, které se často nacházejí mimo běžné chápání čistě technických možností, jež budou mít na rozvoj vyhledávání informací a dat na internetu nemalý vliv. Souběžně s tím je možné vidět zásadní legislativní aktivity, které se jednak snaží regulovat dostupnost nelegálního obsahu, ale současně také omezovat přístup ke svobodným informacím, jak se to děje v totalitních státech.

Klíčovou otázkou pro demokratický vývoj ve společnosti je přitom nalezení správné rovnováhy mezi volným a svobodným přístupem k informacím, které umožnily vznik informační revoluce[9] a možná také určité politické posuny v oblasti středomoří či Číny, ale současně omezují soukromí. Tedy hodnotu, která je pro svobodu a demokracii mimořádně důležitá a stěžejní. [10]

# Bibliografie

[1] ČERNÝ, Michal. Stručný úvod do konceptu sémantického desktopu. Inflow: information journal [online]. 2011, roč. 4, č. 12 [cit. 2012-02-23]. Dostupný z WWW: <http://www.inflow.cz/strucny-uvod-do-konceptu-semantickeho-desktopu>. ISSN 1802-9736.

[2] ČERNÝ, Michal. PROTECT IP Act: konec svobodného internetu?. In: Root[online]. Internet Info, 25. 1. 2012 [cit. 2012-02-23]. Dostupné z: <http://www.root.cz/clanky/protect-ip-act-konec-svobodneho-internetu/> ISSN 1212-8309.

[3] ČERNÝ, Michal. SOPA: Skryté Odepření Práv (nejen) Američanům. In: Root[online]. Internet Info, 18. 1. 2012 [cit. 2012-02-23]. Dostupné z: <http://www.root.cz/clanky/acta-slozita-cesta-k-ratifikaci-a-prakticky-dopad/> ISSN 1212-8309.

[4] ČERNÝ, Michal. ACTA: složitá cesta k ratifikaci a praktický dopad. In: Root[online]. Internet Info, 31. 1. 2012 [cit. 2012-02-23]. Dostupné z: <http://www.root.cz/clanky/acta-slozita-cesta-k-ratifikaci-a-prakticky-dopad/> ISSN 1212-8309.

[5] ČERNÝ, Michal. ACTA: kontroly na hranicích i trestné pirátství pro vlastní potřebu. In: Root[online]. Internet Info, 30. 1. 2012 [cit. 2012-02-23]. Dostupné z: <http://www.root.cz/clanky/acta-kontroly-na-hranicich-i-trestne-piratstvi-pro-vlastni-potrebu/> ISSN 1212-8309.

[6] ČERNÁ, Zuzana, ČERNÝ, Michal. Principy vyhledávání informací na internetu. Metodický portál: Články [online]. 06. 12. 2011, [cit. 2012-02-23]. Dostupný z WWW: <http://clanky.rvp.cz/clanek/c/G/14533/PRINCIPY-VYHLEDAVANI-INFORMACI-NA-INTERNETU.html>. ISSN 1802-4785.

[7] ÖQVIST, Karen Lawrence. Virtual shadows: your privacy in the information society. Swindon, UK: BCS, c2009, 202 s. ISBN 19-061-2409-4.

[8] REED, M.G. Anonymous connections and onion routing. Selected Areas in Communications, IEEE Journal on. 1998, č. 4, 482 - 494.

[9] HOLTZMAN, David H. Privacy lost: how technology is endangering your privacy. 1st ed. San Francisco: Jossey-Bass, c2006, 326 s. ISBN 9780787985110 (CLOTH).

[10] SOLOVE, Daniel J, Marc ROTENBERG a Paul M SCHWARTZ. Privacy, information, and technology. New York: Aspen Publishers, c2006, 321 s. ISBN 07-355-6411-6.