



# ataxo

## Lokální vyhledávání v češtině

26.3.2007

Matěj Novák

Aktuální verzi tohoto dokumentu naleznete na adrese

<http://www.ataxo.cz/firma/tiskove-zpravy/0704-relevance-vyhledavani/>

# Obsah

## Lokální vyhledávání

<b>v češtině</b> .....	<b>1</b>
<i>Shrnutí</i> .....	2
<i>Jak lidé vyhledávají?</i> .....	3
<i>Zobrazení výsledků vyhledávání</i> .....	4
<i>Google</i> .....	4
<i>Seznam</i> .....	4
<i>Centrum</i> .....	4
<i>Atlas</i> .....	5
<i>Shrnutí</i> .....	5
<i>Metoda</i> .....	6
<i>Kategorie a jejich bodové ohodnocení</i> .....	6
<i>Poznámky k metodě</i> .....	7
<i>Výsledky výzkumu</i> .....	8
<i>Komentář výsledků</i> .....	8
<i>Závěr</i> .....	9

## Shrnutí

Úkolem této studie je porovnat relevanci výsledků lokálního vyhledávání nejpoužívanějších českých vyhledávačů (*Seznamu, Centra a Atlasu*) a *Googlu*.

Na základě vyhodnocení relevance více než 2 000 výsledků lokálního vyhledávání dochází studie k závěru, že relevance výsledků *Googlu* je u lokálního vyhledávání výrazně lepší než relevance výsledků hledání sledovaných českých vyhledávačů, ačkoliv *Google* nepracuje s českou gramatikou. Hlavní přednosti *Googlu* se ukazují v tom, že velmi zřídka podává výsledky, které jsou nerelevantní, a že zároveň dosahuje nejvyššího procenta výsledků, které přesně odpovídají dotazu uživatele.

---

\*Lokálním vyhledáváním rozumíme situaci, kdy uživatel hledá „něco někde“ – nejčastěji službu v dané lokalitě (hotel v Třeboni, autoservis v Brně atp.).

## Jak lidé vyhledávají?

Při lokálním vyhledávání má uživatel několik možností, jak v češtině formulovat svůj dotaz. První možností je jazykově správná forma s použitím předložky, např. „autoservis v Brně“. Z hlediska vyhledávačů je podstatné, že v této formě není jméno lokality v nominativu. Druhou možností je použít pouze popis služby spolu s názvem lokality v nominativu – v našem případě tedy „autoservis Brno“. Kromě těchto dvou možností by také uživatel mohl pro určení lokality použít adjektivum a hledat „brněnský autoservis“. Tato možnost je ovšem evidentně velmi málo používaná, proto se můžeme zaměřit pouze na první dvě.

Na první pohled se zdá, že druhý typ fráze bude používanější. Vyhodnocení několika set frází, které použili živatele vyhledávače Jyxo při lokálním vyhledávání, nám tuto domněnku potvrdilo. Zjistili jsme, že druhá forma (tedy použití výhradně substantiv v nominativu) je výrazně častější než první, jazykově korektní způsob. (Byla použita v cca. 95 % případů.) Získat relevantní data z [Google Trends](#), který by byl pro podobné srovnání ideální, se nám bohužel nepodařilo, neboť tento nástroj nemá pro češtinu dost dat. Jedinou výjimkou mohou být slovní spojení, v nichž se vyskytuje sloho „hotel“, avšak zde jde vesměs o hledání anglicky mluvících uživatelů.

V naší studii jsme se proto zaměřili na druhou formu dotazů – tedy použití nominativu pro specifikaci lokality („autoservis Brno“).

# Zobrazení výsledků vyhledávání

Než se budeme zabývat relevancí výsledků vyhledávání, podívejme se nejprve na způsob jejich zobrazení v jednotlivých vyhledávačích.

Všechny vyhledávače zobrazují přirozené (neplacené) výsledky vyhledávání jako modré podtržené titulky následované dvěma až třemi řádky popisu menším nezvýrazněným písmem a URL odkazované stránky. Mezi jednotlivými vyhledávači v tomto ohledu nejsou velké rozdíly, liší se pouze délka a forma popisku.

## Google

Google zobrazuje jako popisek buď obsah shrnutí stránky, který je udaný ve značce `meta description`, nebo katalogový popisek, nebo úryvky ze stránky. Popisek je ve většině případů dlouhý 1–2 řádky (maximální délka je přibližně 150 znaků).

Vpravo od přirozených výsledků vyhledávání jsou zobrazeny kontextové reklamy AdWords. V případě výborné shody s vyhledávanou frází a splnění dalších parametrů mohou být až 3 placené odkazy zobrazeny nad výsledky vyhledávání. Reklamy jsou snadno odlišitelné od přirozených výsledků vyhledávání. (Dle některých průzkumů je však uživatelé často nerozlišují).

Při vyhledávání v češtině se na Googlu v drtivé většině případů nezobrazují žádné speciální či vertikální výsledky jako zprávy, encyklopedie, obrázky atd.

Ve výchozím nastavení zobrazuje Google 10 výsledků na stránku, ve výsledcích se mohou objevit maximálně dvě stránky z jedné domény.

## Seznam

Popisky Seznamu jsou delší než u Googlu – 2 až 3 delší řádky (maximální délka je přibližně 260 znaků). Popisek je tvořen úryvkem ze stránky.

Seznam zobrazuje reklamy nad přirozenými výsledky vyhledávání, vpravo od nich a pod nimi. Reklamy jsou snadno odlišitelné od přirozených výsledků vyhledávání. Nad výsledky fulltextového vyhledávání se rovněž mohou objevit odkazy do katalogu, map nebo encyklopedie.

Při výchozím nastavení zobrazuje Seznam 10 přirozených výsledků vyhledávání.

## Centrum

Centrum zobrazuje 2 řádky popisu (přibližně 220 znaků). Popisek je tvořen buď úryvkem ze stránky, nebo je převzat z katalogu.

Pod přirozenými výsledky vyhledávání se zobrazují 3 reklamy PPC systému AdFox, které jsou snadno rozpoznatelné od přirozených výsledků.

## Atlas

Atlas zobrazuje až 3 řádky (přibližně 300 znaků) pokud je popis převzat z katalogu a 1 řádek (přibližně 120 znaků) pokud je popis tvořen úryvkem ze stránky.

Atlas ve výsledcích vyhledávání kombinuje reklamy a placené zápisy s přirozenými výsledky vyhledávání. První reklamní pozice, na nichž může být až 6 inzerátů, jsou odlišeny modrým pozadím a jsou tak snadno rozeznatelné od přirozených výsledků vyhledávání. Pod nimi se objevují další placené zápisy, které nejsou zřetelně odlišeny od přirozených výsledků vyhledávání.

Na stránce se zobrazuje celkem 20 výsledků, nicméně většina z nich (pravděpodobně až 16) může být a většinou také bývá obsazena placenými inzeráty. V případě lokálního vyhledávání jsou tyto inzeráty obvykle zcela neúčinné, neboť odpovídají buď pouze vyhledávané službě nebo pouze vyhledávané lokalitě, zřídka však oběma. Při hodnocení relevance jsme se zaměřili pouze na přirozené výsledky vyhledávání, nicméně **pro reálné uživatele je zakoušená relevance výsledků výrazně horší**, neboť přirozené výsledky jsou pro uživatele v podstatě nerozlišitelné od placených, které jsou velmi často nerelevantní.

## Shrnutí

Zobrazení výsledků vyhledávání se liší vesměs jen v detailech. Pouze u Atlasu jsou přirozené výsledky vyhledávání pro uživatele na první pohled nerozlišitelné od placených odkazů, což velmi silně poškozuje reálnou použitelnost této služby.

# Metoda

Objektivně hodnotit relevanci výsledků vyhledávání je nepochybně velmi obtížný úkol, neboť požadavky uživatelů se mohou velmi lišit, hledané fráze mohou být mnohoznačné, neexistují objektivní měřítka informační kvality obsahu atd. Posuzování kvality výsledků lokálního vyhledávání je nicméně výrazně snazší, neboť víme, že hledáme „něco někde“ a jsme schopni bez větších obtíží říci, zda jsme to našli, či nikoliv. Při lokálním vyhledávání uživatel očekává, že vyhledávač zobrazí odkazy na stránky, které nabízejí hledanou službu v dané lokalitě, nebo obsahují seznam odkazů na takové stránky. Pokud např. zadá vyhledávací dotaz „kadeřnictví Ostrava“, pravděpodobně ocení buď webové stránky konkrétního kadeřnictví v Ostravě, nebo seznam odkazů, případně kontaktů na kadeřníky v Ostravě.

V rámci naší studie jsme proto požádali 4 uživatele, aby navštívili stránky, které vyhledávače našli při zadání 15 dotazů. Uživatelé vyhodnotili obsah stránek a přiřadili jim kategorie z níže uvedeného seznamu. Každé kategorii jsme přiřadili bodové ohodnocení podle toho, jak dobře uspokojuje požadavek uživatele, totiž najít danou službu v dané lokalitě. Na základě přiřazených kategorií a jejich bodových ohodnocení jsme následně provedli vyhodnocení relevance.

## Kategorie a jejich bodové ohodnocení

**1 – Odpovídající stránka.** Odkaz ve výsledcích vyhledávání přivede uživatele přímo ke službě, kterou hledá, v odpovídající lokalitě. Takový výsledek považujeme za optimální a proto jej hodnotíme 1 bodem.

**2 – Odkaz na odpovídající stránku.** Uživatel se z výsledků vyhledává dostane na stránku, která obsahuje odkaz na jeho „cílovou stránku“ a případně další informace. Uživatel však pravděpodobně bude chtít navštívit i cílovou stránku kvůli získání dalších informací. Za typický příklad stránky tohoto typu považujeme katalogový záznam o firmě. tento typ výsledku považujeme za užitečný, ač jde z hlediska uživatele pravděpodobně o zbytečný krok navíc. Proto tomuto typu stránek přiřazujeme 0.7 bodu.

**3 – Částečně odpovídající stránka.** Uživatel se z vyhledávače dostane na stránku, která neposkytuje hledané informace, avšak není zcela irelevantní. Do této kategorie patří např. stránky, které nabízejí odpovídající službu, avšak v neodpovídající lokalitě. Vzhledem k tomu, že stránka tohoto typu není pro uživatele příliš přínosná, přiřazujeme jí 0.3 bodu.

**4 – Seznam odpovídajících stránek.** Uživatel se dostane na stránku, která obsahuje seznam služeb, které hledá, v odpovídající lokalitě. Tento typ stránky rovněž považujeme za optimální výsledek, neboť dává uživateli přehled a možnost volby, a proto ji hodnotíme 1 bodem.

**5 – Částečně odpovídající seznam.** Výsledek odkazuje např. na stránku se seznamem odpovídajících služeb, ale neodpovídá lokalita – seznam obsahuje jen malé množství odkazů na služby v požadované lokalitě. Najít požadované informace vyžaduje určité úsilí, ale výsledek není zcela neúčinný, proto tento typ stránky hodnotíme 0.5 bodu.

**6 – Irelevantní stránka.** Stránka je z hlediska záměru uživatele zcela nerelevantní a proto ji hodnotíme 0 body.

## Poznámky k metodě

- Výsledky vyhledávání byly získány pomocí standardního prohlížeče ve výchozím nastavení. Na Google bylo zapnuté české rozhraní a hledání nebylo omezeno na stránky v českém jazyce. Dotazy byly pokládány z IP adresy lokalizované v Praze v polovině března 2007. Do úvahy jsme brali prvních 10 výsledků přirozeného vyhledávání z první stránky výsledků.
- Kvůli zajištění objektivity dostali testeři k dispozici pouze seznam URL, která vyhledávače vrátili při zadání daného klíčového slova, bez informace o tom, ze kterého vyhledávače dané URL pochází. Testeři rovněž neznali bodové hodnocení jednotlivých kategorií.
- Pro testování jsme zvolili 5 klíčových slov, u nichž lokalizace hraje důležitou roli. Jako reprezentanty hledání firmy, která nabízí danou službu v okolí jsme zvolili „kadeřnictví“ a „autobazar“. „Nemocnice“ slouží jako příklad veřejné služby, u níž je lokalita rovněž velmi důležitá. Velká část lokálního vyhledávání se týká cestování, proto jsme jako příklad z této kategorie zvolili „ubytování“. Klíčové slovo „byty“ jsme vybrali jako reprezentanta hledání, které sice nepojmenovává danou službu, ale je rovněž typické pro lokální vyhledávání.
- U každé služby jsme použili 3 jména měst, abychom předešli nahodilým zkreslením nevypovídajícími výsledky a abychom zvýšili objektivitu dosažených výsledků. Ve všech případech jsme použili větší města v kombinaci s menšími, abychom zjistili, jak se vyhledávače vyrovnávají s různou „konkurenčností“ v daném segmentu.
- Výsledkem hodnocení bylo 2023 hodnocení výsledků vyhledávání. 14 z nich bylo neplatných – tester buď opomenul vyplnit kategorii, nebo přiřadil dvěma kategoriím 1 bod. (4 vyhledávače × 15 klíčových slov × 4 testeři × 10 výsledků z jednoho vyhledávače – odkazy, které vrátil víc než jeden vyhledávač a byly tudíž zahrnuty jen jednou). Tento počet hodnocených stránek považujeme za dostatečný základ pro objektivní zhodnocení relevance výsledků.
- Je samozřejmě možné namítnout, že by bylo možné stanovit jiné bodové ohodnocení kategorií, případně jinou kategorizaci a dospět tím k odlišným výsledkům, nicméně naše experimentování ukázalo, že rozumně odůvodnitelné změny vah nevedou k zásadním rozdílům ve výsledném hodnocení.

## Výsledky výzkumu

Shrnutí výsledků podává následující tabulka. Úplné výsledky jsou k dispozici na stránce <http://www.ataxo.cz/files/se-relevance/results.html> (v angličtině).<sup>\*</sup> Celkové skóre je průměrem bodových hodnot přiřazených konkrétním výsledkům na základě výše popsané kategorizace.

Search Engine	1	2	3	4	5	6	Celkové skóre
<b>Atlas</b>	25.0%	9.7%	7.9%	7.6%	15.2%	34.7%	<b>0.50</b>
<b>Centrum</b>	29.3%	9.6%	8.3%	9.8%	17.2%	25.8%	<b>0.57</b>
<b>Google</b>	32.2%	9.5%	6.1%	19.4%	14.0%	18.7%	<b>0.68</b>
<b>Seznam</b>	26.8%	10%	6.7%	10.5%	14.9%	31.1%	<b>0.54</b>

*Z tabulky je vidět, že skóre českých vyhledávačů se příliš neliší, avšak Google dosáhl výrazně menšího počtu nerelevantních výsledků (6) než ostatní vyhledávače, našel velké množství dobře zacílených seznamů (4) a našel také nejvyšší počet stránek, které přesně odpovídají zadání uživatele (1) – ačkoliv zde není rozdíl příliš velký. Díky všem těmto faktorům získal Google náskok před ostatními vyhledávači a v celkovém hodnocení je výrazně předčil.*

### Komentář výsledků

- Mezi některými zkoumanými frázemi jsou velké rozdíly v průměrné dosahované relevanci. V některých případech je na internetu mnoho stránek, které jsou pro danou frázi relevantní a všechny vyhledávače vracely téměř perfektní výsledky. Mezi takové fráze patří např. „ubytování Plzeň“. V opačných případech, jako je „kadeřnictví Česká Třebová“, neexistují téměř žádné stránky, které by přesně odpovídaly přání uživatele najít kadeřníka v uvedeném městě, neboť místní kadeřnictví ještě nemají webové stránky. V těchto případech Google dosahoval výrazně lepších výsledků než ostatní vyhledávače, neboť našel alespoň stránky, na nichž byl uveden kontakt na místní kadeřnictví. Ostatní vyhledávače často vracely stránky, na nichž byla někde v textu (často obsáhlém) zmínka o kadeřnictví a na jiném místě jméno města, avšak bez významového propojení.
- Na možnou námitku, že případ zvažovaný v předchozím bodu není příliš typický může spíše zkreslit celkové výsledky, můžeme odpovědět, že i když tyto sporné případy vyřadíme, výsledky se výrazně nezmění. Výsledky bez započtení těchto problematických frází jsou k dispozici na adrese <http://www.ataxo.cz/files/se-relevance/results-1.html>. Kromě toho považujeme schopnost Googlu najít relevantní údaje i tam, kde jde o nalezení virtuální jehly v kupce sena, za jednoznačně užitečnou vlastnost. (Ačkoliv v tomto ohledu samozřejmě nechceme dělat ukvapené závěry na základě hodnocení vyhledávání dvou frází.)
- U některých oblastí naopak vyhledávače často zobrazily několik stránek, které přinášely informace o téže službě (typicky hotelu) bez výrazné přidané hodnoty. Bylo by zajímavé sledovat, jak se jednotlivé vyhledávače dokážou vyrovnat s tímto typem podobnosti, která je z uživatelského hlediska v podstatě duplicitou, nicméně toto téma by vydalo na samostatnou studii a vyžadovalo také jinou metodiku.

\* V případě zájmu o úplná zdrojová data hodnocení kontaktujte autora studie.

- Při hledání klíčového slova „byty“ jsme narazili na zajímavý jev: Seznam a Atlas poměrně často vrací stránky, které nabízejí jeden byt, zatímco Google respektuje plurální formu hledaného výrazu. V našem případě je to záhodné chování, protože uživatel pravděpodobně nehledá informace o jednom konkrétním bytu, nýbrž si chce vybrat ze širší nabídky. Skloňování, které Seznam a Atlas používají, je v tomto případě spíše na škodu věci.
- Zaujalo nás, že [úvodní stránka Karlovarské krajské nemocnice](#) se vůbec neobjevila ve výsledcích vyhledávání Atlasu, Centra ani Seznamu při hledání výrazu „nemocnice Karlovy Vary“, ačkoliv tato stránka je s nejvyšší pravděpodobností **ta stránka**, kterou uživatel hledá. Všechny české vyhledávače místo úvodní stránky nabídli [seznam sponzorů nemocnice](#), který pravděpodobně nebude tím, co běžného uživatele na nemocnici nejvíce zajímá. (Google zobrazil úvodní stránku nemocnice na prvním místě výsledků.) Příčinou tohoto chování by mohl být fakt, že text „nemocnice Karlovy Vary“ se na úvodní stránce nemocnice vyskytuje pouze jednou uprostřed textu, zatímco na stránce se sponzory je obsažen přímo v titulku stránky. Zmiňované vyhledávače pravděpodobně přikládají velkou váhu on-page faktorům, což v důsledku způsobuje toto zvláštní chování. (Možnost, že se vyhledávače tímto způsobem vyjadřují k situaci v českém zdravotnictví ponecháváme stranou. Nelze ovšem vyloučit, že jde o promyšlený krok nemocnice vedoucí k přilákání dalších sponzorů.) S podobným jevem jsme se setkali i u fráze „nemocnice Klatovy“, jejíž úvodní stránka je tvořena pomocí rámu a nemá téměř žádný textový obsah.

## Závěr

Na základě výsledků studie můžeme učinit závěr, že ačkoliv Google nepracuje s českým skloňováním a dalšími specifiky češtiny, jeho výsledky vyhledávání jsou v oblasti lokálního vyhledávání o poznání relevantnější než výsledky českých vyhledávačů. Rozdíly mezi ostatními vyhledávači nejsou velké, výsledky Atlasu z uživatelského hlediska nicméně velmi trpí nevhodným smíšením s placenými odkazy.