

# Oponentský posudek diplomové práce

Název DP:      **Systém pro sledování a analýzu změn stránek vybraných webových serverů**  
Diplomant:     **Michal Kaska**

---

## *Obsah práce:*

Předmětem diplomové práce je implementace nástroje pro sledování a analýzu změn na vybraných webových serverech. Autor dekomponuje dosti komplikovanou úlohu na několik drobnějších úloh a snaží se maximálně využít již existujících nástrojů, které popisuje ve druhé kapitole. Ve třetí kapitole následně vyjmenovává specifické požadavky na vytvářený systém. Ve čtvrté kapitole popisuje obecný návrh systému a základní třídy použité pro realizaci stahování, perzistence a také popisuje, jak budou využívány aplikace třetích stran. V páté kapitole je popsána struktura aplikace, v šesté kapitole jsou ukázky formulářů a uživatelská dokumentace.

## *Hodnocení:*

Textová část práce je dobře strukturovaná, autor uvádí čtenáře do kontextu a popisuje jednotlivé kroky a motivaci pro vytváření takového systému. Nicméně, místy chybí hlubší analýza nebo vysvětlení pro aplikaci celkem zásadních problémů (např. jak se mapuje datový model na objektový). Navržená aplikace působí dojmem beta-verze, tj., byly rozchozeny základní scénáře, ale pro reálné použití je potřeba ještě hodně práce.

V práci nejsou vůbec popsány problémy s nasazováním externích aplikací, po přečtení to vypadá, že se jen jednoduše napasovaly škatulky na sebe a je to hotové. Nicméně, integrace různých systémů je netriviální, je potřeba popsat rozhraní, popř. možné chyby a jejich "záplatování". Bez popisu těchto problémů v práci dojde k tomu, že jednou prošlapaná cesta opět "zaroste".

Při analýze obsahu stránek se autor zmiňuje pouze o HTML značkách, přitom odkazy na jiné zdroje se mohou vyskytovat i v CSS nebo ve skriptovacím jazyce (JavaScript, VBScript, ...). Vůbec není zmíněno, jak se zpracují stránky, jejichž části se generují pomocí skriptu (např. datum, menu, atp...). Metoda LCS je příliš triviální na automatickou analýzu změn. Spíše jen zvýrazní pro uživatele změny, které si pak vyhodnotí sám. Tak jak je to nyní navrženo, budou mít neustále "různé" verze všechny stránky, u kterých je použito aktuální datum. Což výrazně zkresluje statistiky.

Kapitola s experimenty je svojí stručností (pouze jedna strana!) asi největší slabinou textové části. Takový nástroj si přímo zaslouhuje hromadu zajímavých experimentů a jejich prezentaci v diplomové práci - např. detekovat jak velká část českého webu je dosažitelná z nejznámějších portálů? nebo by se dalo vysledovat, jak často se jednotlivé portály mění, atp... Tato část je hodně podceňena, přitom je velice zajímavá pro čtenáře.

*Podrobnější připomínky, poznámky:*

- 1) Dokument obsahuje na hodně místech překlepy, např.

strana	úryvek z textu, chyba zvýrazněna barvou
8	„pasování“ namísto „parsování“
10	„nelézt“ namísto „nalézt“
12	„nepodařilo mu“ namísto „nepodařilo se mu“
...	

- 2) K diagramům chybí slovní komentáře v textu - např. diagram na straně 28 není nikde popsán.
- 3) Obrázky na straně 68 jsou bez popisků, prostřední obrázek obsahuje červeně zvýrazněný JavaScript!?
- 4) Proč je třída WatchFile vázána na konkrétní dvě verze? Nebylo by lepší porovnávat aktuální verzi z Webu s libovolnou již staženou verzí dané stránky?

*Závěr:*

Práce svým průměrným rozsahem dostatečně pokrývá všechny body zadání, autor se v ní tvůrčím způsobem snaží o řešení opravdu netriviálního problému. Práci doporučuji k obhajobě.

V Praze dne 13. května 2010

