

Vyhledávání dat v souboru ve formátu Open Document Spreadsheet (Java) Tým B

Závěrečná zpráva

Vít Novotný

V rámci projektu jsem na počátku nastudoval strukturu formátu Open Document a navrhnul celkovou strukturu výsledného kódu. Samostatně jsem implementoval uživatelské grafické rozhraní obsažené v balíku `gui` včetně jeho české a anglické lokalizace a podílel jsem se na implementaci zbývajících dvou balíků `common` a `cli` implementující hlavní výkonný kód a konzolové rozhraní. V rámci balíku `common` jsem nejen pomáhal s přípravou využitého XML schématu a XSLT transformace, ale i přímo rozšiřoval Javovou implementaci. V rámci balíku `cli` jsem pouze zprostředkoval code reviews.

Zároveň jsem sepsal přehledovou dokumentaci k balíkům `common` a `gui` včetně grafických podkladů. Tuto dokumentaci jsem uveřejnil ve wiki spojené s Gitovým repozitářem projektu v rámci služby GitHub. Dále jsem připravil detailní dokumentaci, která je dostupná formou JavaDocu v rámci zdrojových souborů třídy. Rovněž jsem zodpovědný za vypracování textu licenčního ujednání a zabalení výsledného kódu do javových archivů pro konzolovou a grafickou verzi programu.

Tímto jsem shrnul svůj přínos projektu. Abych vyčerpал požadovaný prostor, popíšu nyní hlavní narativ vývoje. Náš program umožňuje full-textově vyhledávat v sešitech formátu Open Document. Prvotní myšlenkou bylo deserializovat vstupní dokument na javové objekty a následné prohledávání implementovat bez dalšího využití XML technologií. Shodli jsme se, že takové řešení by bylo v rámci tohoto předmětu nedostačující.

V druhotné implementaci jsme se tedy vydali jiným směrem. Vstupní dokument je nejprve zkontrolován s ohledem na strukturu a v dobré víře v jeho validitu je předán XSLT procesoru, který dokument zredukuje na zjednodušený formát vhodný pro vyhledávání.

Toto řešení je robustní v tom smyslu, že validita vstupního dokumentu je podmínkou dostačující, ale nikoliv nutnou, pro úspěšnou transformaci a lze tedy zpracovávat i dokumenty v rozšířeném formátu, které mohou nad rámec schématu obsahovat libovolné dodatečné elementy a atributy, což znemožňuje jejich strojovou validaci.

Výsledný dokument v zjednodušeném formátu je validován vůči příslušnému XML schématu a následné vyhledávání probíhá formou XPath výrazů. Shodli jsme se, že toto řešení bylo lepším testem znalostí nabytých v rámci tohoto předmětu.