Záverečná správa projektu

Filip Daniel Fedin

Úvod

Mojou úlohou v tíme bolo vyextrahovať dáta z pdf dokumentu do vhodného formátu, ktorý by následne ďalší členovia tímu vedeli dobre, ľahko a jednoducho spracovať do XML súboru.

Vypracovanie

Vytvoril som classu Role, s attribútmi podľa XSD schémy ktorú vytvorila Michaela. Z tohto objektu by sa potom ľahko generoval XML súbor. Prišiel som na dva prístupy k tomu ako by sa dali tie dáta získať. Nástroj tabula - ktorý dokázal získať data z PDF tabuliek, a knižnicu Apache PDFBox. Tá by síce tiež dokázala spracovať tie tabulky, ale bolo by to oveľa zložitejšie, a celkom jednoducho dokázala vytiahnúť čistý text z PDF dokumentu. Problém bol ale v tom, že jednotlivé Deliverables z dokumentu boli v tabulke usporiadané tak, že neboli nijako oddelené. To znamená, že nemal ako zistiť ktorá kde patrí. Rozhodol som sa teda použiť tabulu, lenže tento nástroj nefungoval úplne korektne, dáta ktoré získal boli celkom rozhádzáne a nekonzistentné. Vzhľadom na to, že dokumenty boli chybné, občas niekde chýbala čiara ohraničujúca tabuľku. Takáto chyba dokázala úplne rozhádzať miestami dáta získane tabulou. Rozhodol som sa teda, že natvrdo vložím do programu počty jednotlivých deliverables, a potom som dokázal sparsovať dokument relatívne bezproblémov, dôležité data boli oddelené kľúčovými slovami, alebo inými deliacimi znakmi ako je napr. odrážka.

Záver

Podarilo sa mi teda vyextrahovať dáta z dokumentu, a moja časť programu generovala list Role objektov, pre každú jednu rolu z dokumentu