Návrh softwarového projektu (předmět NPRG023) pro šk. rok 2011/2012

Název: Digitalizace a zpřístupnění vzpomínek pamětníků z archívu Konfederace politických vězňů (DIGIPAM)

Vedoucí: Jan Hajič (ÚFAL MFF UK)
Konzultanti: Jakub Mlynář (CVHM MFF UK), Milan Fučík (ÚFAL MFF UK)

Cíl projektu:

Digitalizovat videonahrávky z archívu KPV a zpřístupnit je zájemcům prostřednictvím terminálů v CVHM jako doplněk k archívu USC-VHI (holocaust).

Náplň projektu:

- design, specifikace a implementace databáze nahrávek, včetně ukládání transkripce, popisu metadat, klíčových slov přidělovaných krátkým segmentům nahrávek, a synchronizace všech typů dat na časové ose (video, audio, klíčová slova, segmenty, plný transkript)
- vytvoření zpřístupňujícího a vyhledávacího systému (web access / search, ukládání vlastních projektů uživatelů, tj. včetně registrace, logování, prac. prostor pro uživatele apod.)
- naplnění databáze pro testování software
    - digitalizace a transformace videa a audia nahrávek do minimálního, ale kvalitního formátu pro streamování přes lokální síť
    - rozpoznání audia nahrávek (transkripce) (pomocí externího software, ZČU Plzeň)
    - přidělení klíčových slov segmentům nahrávek (vlastní design klasifikátoru, klasifikátor ne nutně vlastní – externí OSS), zajištění alespoň 70% kvality (experimenty, vyhodnocení)

Zásady pro vypracování:

Vyhledávací aplikace musí fungovat pod hlavními browsery (IE, Chrome, Mozilla, Safari, Opera). Databáze musí být založena na OSS. (Persistentní) pracovní prostor (databáze) pro uživatele musí být oddělena od vlastních nahrávek. Přístupový systém musí umožňovat „superuser“ přístup z technického i admin. hlediska. Databáze musí být zálohovatelná (garance 24/7 přístupu není nutná); výběr OS je na projektovém týmu. Transkripce provedené automaticky je třeba vyhodnotit a případně provést manuální korekce (zajistí ÚFAL mimo projekt, bude-li nutno; neprovedení ale neohrožuje projekt, neboť se mohou pouze zhoršit výsledky vyhledávání). Fulltextová komponenta vyhledávání bude využívat lingvistické nástroje ÚFAL (min. lematizaci), specifické požadavky bude třeba řešit v rámci projektu. Tezaurus pro indexování klíčovými slovy dodá ÚFAL v minimální versi nutné pro otestování SW komponent projektu.

Počet členů týmu: min. 4, max. 5

Zajištění projektu:

Finální HW zajistí ÚFAL (M. Fučík). Vývoj může probíhat zcela nezávisle. Nahrávky poskytne již při zahájení projektu CVHM (J. Mlynář). Licenční ujednání bude řešit ÚFAL a CVHM (netýká se účastníků projektu).

Kontakt: hajic@ufal.mff.cuni.cz; pro info o CVHM: http://malach-centrum.cz