# V. Popis projektu

*(Text max. 10 běžných stran formátu A4)*

**Povinná osnova popisu projektu (nutno vyplnit všechny body)**

**1. Vymezit konkrétní cíl(e) projektu v souladu s jedním či více specifickými cíli globálního cíle programu a způsob jejich naplnění.**

Hlavním cílem projektu je vývoj softwarových nástrojů, které pomohou zpřístupnit rozsáhlý archív dokumentů a nahrávek orální historie pro badatele, pedagogy, dokumentaristy i širokou veřejnost a také zjednoduší a zefektivní archivaci takových nahrávek. Dojde tak k naplnění specifického cíle 1.1, konkrétně bodu (d) - vytváření metod dokumentace a prezentace paměťové kultury národa.

Jádrem projektu je výzkum metod a jejich následná implementace do softwarových nástrojů umožňujících snazší zpracování a zpřístupnění rozsáhlého množství záznamů a dokumentů textového a zvukového formátu. V rámci projektu bude zpracováno minimálně 50 000 textových dokumentů a 1000 hodin záznamu audionahrávek rozhovorů a výpovědí, které vznikly v rámci dokumentační činnosti Ústavu pro studium totalitních režimů (dále ÚSTR) v rozmezí let 2008 až 2015. Jde o bilanční rozhovory s pamětníky totalitních režimů v Československu. K jednotlivým rozhovorům lze najít a přiřadit příslušné kopie dokumentů a fotografií z domácích archivů a dalších zdrojů (v případě pamětníků Gulagu například kopie spisů NKVD pracně dohledávaných v bývalých zemích SSSR). Tato unikátní sbírka historických pramenů je využívána dle aktuálních projektů či výstupů ÚSTR (odborné studie, dokumentární pořady, výstavy, monografie atd.). S nárůstem objemu sbírky i růstem poptávky po zpřístupnění materiálu ze strany externích historiků, dokumentaristů a široké veřejnosti má ÚSTR zájem na zprovoznění takového řešení, které bude nejen efektivní pro provozovatele (tj. mimo jiné umožní relativně snadnou, přehlednou a také časově a finančně efektivní archivaci), ale i pro badatele či zájemce z řad veřejnosti veřejnost, kteří archiv budou chtít využívat.

Integrovaný archiv nahrávek, dokumentů a fotografií přístupný online a prohledávatelný podle různých aspektů (konkrétní obsah nahrávek, jméno a ostatní životopisné údaje pamětníka, časové období, ke kterému se dokument vztahuje, apod.) by tedy práci uživatelů výrazně zefektivinil a zároveň by zpřístupnul informace v něm obsažené mnohem širšímu okruhu uživatelů.

V současné době jsou nahrávky společně s protokolem o natáčení, souvisejícími dokumenty a fotografiemi deponovány na interním uložišti ÚSTR a zpřístupňovány badatelům na DVD, případně zasílány přes digitální uložiště. Pouze cca 160 nahrávek se dosud podařilo opatřit ručním přepisem (transkripcí) obsahu rozhovoru. U ostatních nahrávek musí badatelé procházet celý videozáznam. Snahy o zefektivnění archivace a zpřístupnění této rozsáhlé a rozmanité sbírky pramenů v minulosti narážely na technické, finanční a personální možnosti ÚSTR. Přes tyto nedostatky je sbírka využívána odbornými pracovníky nejen z českých zemí, ale i z ostatních zemí Evropy a Spojených států. Historické prameny v ní obsažené, byly využity v řadě odborných i populárně-naučných publikacích. Navrhované zpracování

sbírky výrazně zhodnotí její potenciál zejména z těchto důvodů:

Tématické, slovní, frázové i fonetické vyhledávání umožní badateli orientaci přímo v nahrávce rozhovoru. Odpadne tím nutnost procházet celé nahrávky, případně nákladně pořízené ruční přepisy (transkripty).

Poloautomatické zpracování naskenovaných dokumentů metodami strojového vidění a navazujícím modulem zpracování přirozeného jazyka umožní rychlejší a efektivnější anotaci jednotlivých dokumentů pomocí metadat a také značně zjednodušší případné přiřazení korespondence a vazeb mezi jednotlivými dokumenty, které mohou být i různé mediální povahy (text a audio, eventuelně fotografie). Díky tomu bude mít badatel k dispozici mnohem širší kontext zkoumané nahrávky.

Těžištěm projektu bude především:

vývoj systému pro komplexní zpracování materiálu ÚSTR zahrnující automatické či poloautomatické zpracování a třídění naskenovaných dokumentů, automatický přepis zvukových nahrávek a jejich následný převod do formy vhodné pro vyhledávání klíčových slov či frází a to jak ve slovní, tak i fonetické podobě. V oblasti zpracování audionahrávek bude využito rozsáhlé zkušenosti řešitelského týmu získaných při řešení projektu AMALACH, NAKI (DF12P01OVV022, 2012 - 2015); z důvodu použití jiných mikrofonů a tématu rozhovoru bude nutno vyvinout nový akustický model a zpracovat nový specifický slovník a jazykový model systému. V oblasti zpracování textových dokumentů bude využito know-how týmu z oblasti zpracování digitalizovaného obrazu. Další část výzkumu pak bude věnována detekci témat. Tento krok bude nutný k nalezení odpovídajících odkazů z daného dokumentu (nahrávky) na další související dokumenty jiné povahy (např. textové).

vývoj softwarového repozitáře pro bezpečné dlouhodobé uložení původních digitálních materiálů (nahrávek, dokumentů) i nově v projektu vytvořených dat z těchto původních materiálů odvozených. Jako základ pro tento repozitář bude použit open source systém pro správu digitálních knihoven LINDAT/CLARIN Repository (https://github.com/ufal/lindat-repository), čímž bude zaručena mimo jiné neměnnost uložených záznamů a jejich citovatelnost jako digitálních objektů pomocí permanentních odkazů Handle systému (http://handle.net). Repozitářový systém LINDAT/CLARIN je založen na open source systému Dspace, který byl v projektu velké infrastruktury LINDAT /CLARIN (LM 2010013) adaptován pro správu jak otevřených dat, tak i dat s přístupem částečně nebo zcela omezeným v závislosti na licenci pro daná data použité. Systém bude podstatně adaptován pro potřeby lepšího zpřístupnění audiovizuálních dat provázaných s dalšími dokumenty různých modalit (převážně textový a obrazový materiál).

Výsledný produkt v podobě komplexního řešení archivace rozmanité sbírky historických pramenů a vyhledávání v ní, bude možné aplikovat u obdobných projektů, potýkajících se s totožnými problémy. Z expertních konzultací s partnerskými institucemi (Post bellum, Stopy paměti) je zřejmý zájem o jeho využití (viz přiložený dopis obecně prospěšné společnosti Post Bellum). Podle konkrétního postupu a kapacit projektu bude případně možno začlenit k rozpoznávání češtiny v omezeném rozsahu i slovenštinu, protože archiv spadá převážně do období Československa a mnoho relevantních záznamů je proto i v

druhém úředním jazyce té doby.

2. **Uvést zda byl nebo je předmět výzkumu v minulosti řešen v rámci jiné výzkumné aktivity podporované z veřejných zdrojů a pokud ano, uvést její identifikaci.**

Úloha komplexního zpracování dat předmětného formátu a rozsahu nebyla dosud řešena. Systém vyhledávání klíčových slov a frází ve zvukových záznamech jiné oblasti - přeživších svědků holokaustu - je vyvíjen v právě končícím projektu AMALACH, financovaného v rámci programu NAKI (DF12P01OVV022, 2012 - 2015). Systém využívá statistický akustický model vytvořený (natrénovaný) speciálně na audionahrávkách svědků holokaustu pořízených jedním typem mikrofonu a jazykový model vytvořený z přepsaných 100 hodin dat těchto nahrávek. Podobné problematice se věnoval také projekt NAKI "Zpřístupnění archivu Českého rozhlasu pro sofistikované vyhledávání" (DF11P01OVV013 - 2011 - 2013). Jak však bude uvedeno dále, vzhledem k odlišnému charakteru nahrávek, systém vyhledávání ve výpovědích svědků holokaustu ani systém pro zpřístupnění archivu Českého rozhlasu nelze v předkládaném projektu přímo použít. Budou ale plně využity všechny získané zkušenosti, know-how i softwarové nástroje, které má řešitelský tým k dispozici z předchozího projektu.

3. **Rozbor stavu řešení problému v ČR a v zahraničí s odpovídajícími citacemi v odborné literatuře.**

Úloha automatického zpracování velkých souborů zpracování komplexních multimodálních dat se v poslední dekádě dostává stále více do centra pozornosti vědeckých týmů, což souvisí nejen s postupujícím pokrokem v oblasti metod strojového učení a nárůstem výpočetních kapacit, ale především s rostoucím objemem shromažďovaných, nahrávaných a archivovaných textových a řečových dat.

V České republice má obor automatického zpracování mluvené řeči dlouhou tradici a úlohou rozpoznávání souvislé mluvené řeči za účelem následného vyhledávání v indexovaných nahrávkách se kromě pracoviště hlavního řešitele zabývají především pracoviště Technické univerzity v Liberci a Vysokého učení technického v Brně. Současné systémy umožňují slovní [1], popř. i fonetické vyhledávání [2] [3]. . Zmíněné fonetické vyhledávání umožňuje najít i slova, která se nevyskytují ve slovníku rozpoznávacího systému nebo si uživatel není jist jejich přesným zápisem (což se ve vzpomínkách pamětníků bude stávat poměrně často - připomeňme, že se podstatná část popisovaných událostí odehrává v zemích bývalého SSSR). V předkládaném projektu bude třeba (pro dosažení co největší úspěšnosti přepisu) vytvořit úloze specifický akustický model (pořízené nahrávky jsou zaznamenávány různými nahrávacími zařízeními (různými druhy audio záznamníků i mikrofonů) a jazykový model (téma výpovědí o životě v totalitním režimu).

Fonetické vyhledávání v řečových datech se zaměřuje především na takové úlohy, kde je potřeba vyvinout metody prohledávání v jazycích, pro které jsou k dispozici jen velmi omezená nebo žádná anotovaná data [4,5]. Omezené zdroje trénovacích dat však samozřejmě způsobují nižší úspěšnost vyhledávání, proto naší snahou bude vyvinout

jazykový model pokrývající co nejvíce slov ve zpracovávaných nahrávkách a fonetické vyhledávání použít pro detekci slov mimo slovník rozpoznávacího systému.

V oblasti automatického rozpoznávání znaků (OCR) je třeba rozlišovat mezi rozpoznáváním znaků z homogenních (tj. většinou strojově psaných) dokumentů, kde v ideálním případě všechny znaky náleží do jednoho nebo několika málo fontů, a rozpoznáváním znaků z dokumentů nehomogenních, obsahujících např. text napsaný různými typy psacích strojů či dokonce ručně.

Zatímco v prvním případě lze použít přístup založen na porovnávání detekovaných znaků s uloženými vzory [6], nehomogenní dokumenty (které budou tvořit v našem případě drtivou většinu) se zpracovávají přístupem založeným na extrakci obrazových příznaků, které se pomocí metod strojového učení rozpoznají jako některý známý znak [7], [8]. Samotný obrázek textu je nutné předzpracovat tak, aby byl vhodný pro rozpoznávání. Jedná se o odstranění šumu, zarovnání, jasové transformace, detekci oblastí zájmu [9] a jiné. Dále je nutné výstup rozpoznávače dále upravit. Pro tento účel se používají metody zpracování přirozeného jazyka. Mezi elementární metody spadá využití slovníků a jazykových modelů (podobných těm, které jsou využívány v úloze rozpoznávání řeči), ale lze použít i znalosti vyšší úrovně (předpokládané rozvržení dokumentu apod.).

Úlohou detekce témat se členové řešitelského týmu také dlouhodobě zabývají a to jak v doméně psaných textů [10], tak i v doméně detekce tématu z rozpoznaných řečových nahrávek.

[1] Nouza, J. et al: Speech-To-Text Technology to Transcribe and Disclose 100,000+ Hours of Bilingual Documents from Historical Czech and Czechoslovak Radio Archive, In: Proceedings of INTERSPEECH 2014, 2014, pp. 964-968.

[2] Karakos, D. and Schwartz, R.: Subword and phonetic search for detecting out-of-vocabulary keywords: In Proceedings of INTERSPEECH 2014, 2014, pp. 2469-2473.

[3 Psutka, J. et al: System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. EURASIP Journal on Audio, Speech and Music Processing, 2011(10), 2011, pp. 1-19.

[4] Hsiao, R., Ng. T., Grézl F., Karakos D., Tsakalidis S., Nguyen L. a Schwartz R.: Discriminative Semi-supervised Training for Keyword Search in Low Resource Languages. In: Proceedings of ASRU 2013. 2013, pp. 440-445.

[5] Trmal, J et al: A keyword search system using open source software, In: Proceedings of SLT 2014, 2014.

[6] Lucas, S.M., Tams, A.C., Cho, S.J., Ryu, S., Downton, A.C.: Robust Word Recognition for Museum Archive Card Indexing. In: Proceedings of Sixth International Conference on

Document Analysis and Recognition, 2001, pp. 144-148.

[7] Farhad, M.M., Nafiul Hossain, S.M., Khan, A.S., Islam, A.: An Efficient Optical Character Recognition Algorithm Using Artificial Neural Network by Curvature Properties of Characters. In: Proceedings of ICIEV 2014, 2014, pp. 1-5.

[8] Shastry, S., Gunasheela, G., Dutt, T., Vinay, D.S., Rupanagudi, S.R.: "i" — A novel algorithm for optical character recognition (OCR), In: Proceedings of Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) 2013, 2013, pp.389-393.

[9] Steinke, K.-H.: Improvement of Omnipage 18's efficiency, In: Proceedings of CSAE 2012, 2012, pp.434-438.

[10] Švec J. et al: General framework for mining, processing and storing large amounts of electronic texts for language modeling purposes. Language Resources and Evaluation, 48(2) 2014, pp. 227-248.

**4. Uvést metodiku řešení projektu.**

Existující systém pro automatické rozpoznání zvukových nahrávek a jejich následný převod do formy vhodné pro vyhledávání klíčových slov či frází bude třeba upravit pro použití ke zpracování nahrávek odlišného charakteru. Dlouhodobé zkušenosti ukazují, že pro tyto účely je nejlepší přepsat část tzv. "cílových nahrávek'' (tj. těch, pro které bude adaptovaný systém používán) tak, aby přepisy byly zarovnány zhruba na úrovni vět. V našem případě je však možné částečně využít již existující přepisy pořízené ÚSTR, které jsou ale přiřazeny k řečovému záznamu pouze na úrovni celých nahrávek. K automatickému přiřazení textového přepisu k nahrávkám bude využit systém vyvinutý pro jinou úlohu, např. úlohu výpovědí svědků holokaustu, Systém musí být doplněn algoritmem detekce míst nesprávného přepisu nahrávky (text neodpovídá audiu), která jsou tudíž nevhodná jako trénovací data. Softwarová implementace systému po doplnění o nezbytné uživatelské rozhraní bude i jedním z výstupů projektu (hlavní výsledek ALIGN - typ R), neboť podobná situace (část dat určená k archivaci je přepsaná, ale přiřazení je příliš "hrubé") nastává v případě zvukových archivů poměrně často. Tímto způsobem tedy nejprve připravíme trénovací data, která budou následně použita pro natrénování či adaptaci akustických, popř. i jazykových modelů.

Nové modely využijeme pro zpracování celé kolekce nahrávek. Výsledkem budou tzv. mřížky, které obsahují nepjpravděpodobnější hypotézy o obsahu zpracovávané promluvy a to jak ve slovní, tak i ve fonetické podobě. Tyto mřížky pak budou uloženy ve formě indexu, která umožňuje efektivní prohledávání. Předpokládáme, že fonetická reprezentace bude využita především pro vyhledání raritních jmen a geografických názvů, které se neobjeví v manuálně přepisované části nahrávek a tudíž nebudou ani obsažena ve slovníku rozpoznávače.

Při plnění repozitáře bude pro budoucí efektivní vyhledávání potřeba označit jednotlivé nahrávky také pomocí metadat, což mohou být například:

- jméno a ostatní životopisné údaje pamětníka či ostatních osob v nahrávce zmíněných
- témata, o kterých se v nahrávce hovoří
- časové období, ke kterému se obsah nahrávky vztahuje
- odkazy na jiné dokumenty (ať už mluvené, textové či obrazové) relevantní k obsahu nahrávky

Předpokládáme, že tento metadatový popis bude vzhledem ke značně heterogenní skladbě dostupných dokumentů prováděn ručně, ale zaměříme se na vývoj inteligentních podpůrných nástrojů, které by práci anotátora co nejvíce usnadnily. Mezi takové podpůrné moduly bude patřit zejména software pro automatickou detekci témat z textu a mluvené řeči (jeho laboratorní verzi již máme k dispozici z předchozích projektů) a software pro automatickou detekci textů z naskenovaných dokumentů. Druhý jmenovaný modul bude založen na automatickém rozpoznávání znaků (OCR) - pro vlastní, "nízkoúrovňové" OCR plánujeme využít již existující volně dostupné nástroje (např. Tesseract), které doplníme nadstavbou využívající jak metody zpracování přirozeného jazyka (jazykové modely, detektory sémantických konceptů), tak i apriorní informace o struktuře zpracovávaných dokumentů.

Pro integraci všech zdrojů použijem repozitářový systém LINDAT/CLARIN, který je založen na open source systému Dspace a byl v projektu velké infrastruktury LINDAT/CLARIN (LM 2010013) adaptován pro správu jak otevřených dat, tak i dat s přístupem částečně nebo zcela omezeným v závislosti na licenci pro daná data použité. Systém bude podstatně adaptován pro potřeby lepšího zpřístupnění audiovizuálních dat provázaných s dalšími dokumenty různých modalit (převážně textový a obrazový materiál). Pro všechny typy uložených dokumentů budou vyvinuta vhodná metadatová schémata a vyhledávání v repozitáři bude optimalizováno na základě těchto metadatových schémat. Každý záznam i vyhledávací systém repozitáře bude také provázán se specializovaným systémem pro tématické, slovní i fonetické vyhledávání v obsahu nahrávek. Tímto způsobem vyvineme hlavní výsledek projektu - HIDOAR, software pro poloautomatické zpracování a zpřístupnění textových a zvukových nahrávek v integrovaném archivu pramenů.

**5. Stručně popsat vybavenost pracoviště – materiální, laboratorní, přístrojové, případně jiné vybavení řešitelského pracoviště nebo pracovišť, přístup k informačním zdrojům potřebným k řešení projektu.**

Pracoviště navrhovatele (ZČU) má k dispozici mimo jiné nové výpočetní stroje vybavené výkonnými GPU kartami pro náročné výpočty, má přístup ke zdrojům Národní Gridové Infrastruktury MetaCentrum, jejíž členství umožňuje přístup k dalším masivním výpočetním prostředkům cloudového typu. Tyto prostředky budou využity k zpracování velkého množství dat pro vývoj a testování statistických modelů systému rozpoznávání řeči, zpracování digitalizovaného obrazu a zpracování přirozeného jazyka. Dostatečné výpočetní zdroje jsou k dispozici také na partnerském pracovišti UK, a to včetně kapacit pro uložení

| |
|---|
| velkého množství multimediálních dat. Obě univerzitní pracoviště také disponují značným množstvím různorodých jazykových dat v elektronické podobě (korpusů), která lze využít pro podporu trénování systémů pro rozpoznávání řeči a další strojové zpracování přirozeného jazyka. Partner ÚSTR je klíčový pro zprostředkování přístupu k řečovým nahrávkám a další různorodým informačním zdrojů, které jsou v centru zájmu předkládaného projektu. |

6. **Specifikovat výsledky projektu (výčet všech očekávaných výsledků).** *Očekávané výsledky musí být rozděleny na výsledky **hlavní a vedlejší**. **Rozdělení jednotlivých druhů výsledků do skupin hlavní a vedlejší výsledky je uvedeno v zadávací dokumentaci, v části 5.4. Očekávané výsledky.** Výsledek musí být specifikován písmenem a textem uvedeným v zadávací dokumentaci (viz tabulka „Pomocné kritérium pro hodnocení poskytovatele z hlediska naplnění indikátorů programu NAKI II"). Povinnou součástí specifikace každého předpokládaného výsledku projektu je:*

| | |
|---|---|
| **písmeno označující druh výsledku** (např. R, G, B atd.) | R (2x), D (9x) |
| **kategorie výsledku**: hlavní/vedlejší (lze uvést v záhlaví pro celou skupinu, pokud od jedné kategorie bude více druhů výsledků a/nebo vícečetné výsledky jednoho druhu výsledku stejné kategorie) | hlavní, vedlejší |
| **předpokládaný název výsledku** | ALIGN, HIDOAR,<br>Článek ve sborníku odborné konference - 9x |
| **krátká charakteristika výsledku** | ALIGN - software pro podporu poloautomatického zarovnání nahrávek s existujícími přepisy pro účely efektivní přípravy dat určených pro trénování akustických a jazykových modelů.<br><br>HIDOAR - software pro poloautomatické zpracování a zpřístupnění textových a zvukových nahrávek v integrovaném archivu pramenů.<br><br>Článek ve sborníku odborné konference - 9x |
| **předpokládaný rok uplatnění výsledku** | ALIGN - 2017,<br>HIDOAR - 2018,<br>články - 2017, 2018, 2019 |
| předpokládaní **budoucí uživatelé výsledku** (tento údaj o uživatelích výsledku nebude uváděn pouze u výsledků publikačních typu B, C, D a J; u ostatních druhů | Obecně prospěšná společnost Post Bellum - portál Paměť národa<br>(doklad o předběžném zájmu |

| | |
|---|---|
| výsledků hlavních i vedlejších pro program NAKI II je nepominutelný). Doklad o zájmu budoucího možného uživatele o navrhovaný výsledek je možné přiložit k přihlášce projektu, pokud jej lze zajistit. | předložen) |
| **dedikace výsledku -** u vedlejších výsledků bude uvedeno, zda bude výsledek dedikován výlučně k projektu NAKI II. Pokud nebude výlučně vázaný na NAKI II (s výjimkou výsledku druhu B - odborná kniha, A - specializovaná veřejně přístupná databáze, kde je tento postup vyloučen - viz ZD), je nutné uvést všechny souvztažné výzkumné aktivity, z kterých bude výsledek rovněž podporován, instituce a autory, kteří se budou na výsledku rovněž spolupodílet. | všechny vedlejší výsledky budou dedikovány výlučně projektu NAKI II |

*V případě, že uchazeč předpokládá více jak jeden výsledek přísl. druhu výsledku, je nutné uvést jejich počet a specifikace u každého z očekávaných výsledků příslušného druhu výsledku.*

*U specifického výsledku pro program NAKI II E - uspořádání výstavy - je nutné dodržet podmínky uvedené v zadávací dokumentaci v části 5.4., včetně zveřejnění publikace typu B (která bude kritickým katalogem výstavy a která musí být v přihlášce projektu jednoznačně jako kritický katalog výstavy označena - v poli krátká charakteristika výsledku). U očekávaných a v přihlášce vymezených výsledků uvést případný mezinárodní přínos. Dále se doporučuje respektovat programem pro daný specifický cíl očekávané druhy výsledků případně další výsledky aplikovaného výzkumu a experimentálního vývoje definované v platné Metodice hodnocení výsledků výzkumných organizací a hodnocení výsledků ukončených programů. Při hodnocení projektu nebude brán zřetel na uvedené očekávané výsledky, které neodpovídají druhům výsledků uvedených ve struktuře RIV15 (např. rukopis, studie, abstrakt apod.).*

***Na závěr bodu 6. bude povinně vyplněna níže uvedená přehledová tabulka počtu předpokládaných výsledků projektu odpovídající komentáři v bodě č. 6.*** *Definice druhů výsledků jsou uvedeny v platném znění Metodiky hodnocení výsledků výzkumných organizací a výsledků ukončených programů.*

| předpokládané výsledky projektu | počet |
|---|---|
| **Hlavní výsledky** | |
| **$F_{uzit}$** - užitný vzor | |
| **$F_{prum}$** - průmyslový vzor | |
| **$G_{prot}$** - prototyp | |
| **$G_{funk}$** - funkční vzorek | |
| **$N_{met}$** - certifikovaná metodika | |
| **$N_{pam}$** - památkový postup | |
| **$N_{map}$** - specializovaná mapa s odborným obsahem | |
| **P** - patent | |
| - "evropský“ patent (EPO), patent USA (USPTO) a Japonska | |
| - český nebo národní patent (s výjimkou patentu USA a Japonska), který je využíván na základě platné licenční smlouvy | |
| - ostatní patenty[7] | |
| **R** - software | 2 |

[7] Český nebo jiný národní patent udělený, doposud nevyužívaný nebo využívaný vlastníkem patentu.

| předpokládané výsledky projektu | počet |
|---|---|
| **$Z_{polop}$** - poloprovoz | |
| **$Z_{tech}$** - ověřená technologie | |
| **$H_{leg}$** - výsledky promítnuté do právních předpisů a norem | |
| **$H_{neleg}$** - výsledky promítnuté do směrnic a předpisů nelegislativní povahy závazných v rámci kompetence příslušného poskytovatele | |
| **E** - uspořádání výstavy - **specifický výsledek programu NAKI**<br>Jedná se o nejméně dva měsíce trvající veřejnou prezentaci kulturních či kulturně historických hodnot s minimální návštěvností 1000 návštěvníků za dobu trvání výstavy, která je výlučně výsledkem výzkumných projektů v rámci Programu aplikovaného výzkumu a vývoje národní a kulturní identity (NAKI), a její součástí je kritický katalog s řádně přiděleným ISBN, jehož obsah prošel recenzním řízením. O případné výnosy ze vstupného musí být poníženy způsobilé náklady projektu. | |
| **Vedlejší výsledky** | |
| **A** - audiovizuální tvorba, elektronické dokumenty | |
| **B** - odborná kniha | |
| **C** - kapitola v odborné knize | |
| **D** - článek ve sborníku (z konference) | 9 |
| **J** - recenzovaný odborný článek | |
| **M** - uspořádání konference | |
| **W** - uspořádání workshopu | |

7. **Poslání a očekávané přínosy projektu ve vazbě na očekávané přínosy programu NAKI** *(část 2.3 programu), včetně zdůvodnění potřebnosti projektu pro naplnění cílů programu NAKI.*

Cílem přeloženého projektu je, jak již bylo zmíněno výše, navrhnout a implementovat softwarové nástroje, které pomohou k efektivnější archivaci a zpřístupnění nehmotného kulturního dědictví (konkrétně vzpomínek pamětníků totalitních režimů a souvisejících historických dokumentů). Tím přispějeme k očekávaným přínosům programu 1. (uchování hmotného i nehmotného kulturního dědictví pro další generace) a 3. (zpřístupnění kulturního dědictví široké obci zájemců a uživatelů). Současný způsob archivace totiž způsobuje, že orientace v datech je velmi obtížná a navíc nejsou jednotlivé datové zdroje (audionahrávky, listinné dokumenty, fotografie) příliš propojeny. Vzhledem k tomu, že projekt plánuje masivně využít nejmodernější IT technologie, přispívá též k očekávanému přínosu 7.k. (výzkum nástrojů a jejich ověření pro systematické a efektivní využívání nových moderních technologií ve společenských, humanitních i technických vědách pro vzdělávání).

8. **Kritické předpoklady dosažení cíle projektu, popis rizik projektu.**

Kromě obecných rizik, která v podobných projektech nastávají vždy (odchody kvalifikovaných pracovníků, apod.) lze pro tento projekt identifikovat tato konkrétní rizika:

- nedostatečná kvalita nahrávek pro natrénování akustického modelu

• nedostatečně přesný výstup z OCR modulu způsobený nízkou kvalitou zpracovávaných dokumentů

První ze zmíněných rizik lze zmírnit či zcela odstranit buď pomocí pokročilých metod zpracování signálu, které pracoviště navrhovatele rutinně používá, nebo využitím dodatečných řečových korpusů, jejichž rozsáhlé portfolio mají obě univerzitní pracoviště k dispozici. Nedostatečnou kvalitu OCR výstupu lze částečně eliminovat pomocí nadstavbových metod zpracování přirozeného jazyka; v krajním případě, jak již bylo řečeno, přiřadíme dokumentu příslušná data mauálně. V oblasti samotné technologie rozpoznávání řeči a následného vyhledávání informací je riziko jen malé, neboť tyto technologie má řešitelský tým dobře zvládnuté.

## 9. Etapy projektu

*Pro každou etapu projektu je nutné uvést (etapy na sebe musí časově a věcně navazovat, popř. se mohou částečně překrývat, ale musí být uvedeny a nesmí být všechny plánovány na celou dobu řešení)*:

a) Číslo, název a cíl etapy

01 - Příprava dat a datových struktur, testy existujících metod

ZČU: Vývoj softwaru ALIGN, testy existujících metod OCR, vývoj metod zpracování přirozeného jazyka pro potřeby archivu

UK: Návrh metadatových schémat repozitáře, testy metod zpracování textu

USTR: Návrh struktury metadat z pohledu badatelů, anotace nahrávek

b) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2016-03-01

c) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2016-12-31

d) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

e) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

R – sw ALIGN

D – článek ve sborníku (plán UK: konference LREC/ACL-EACL/ TLT 2016)

f) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN.

Popisu funkčnosti výsledku typu R společně s licenčními podmínkami pro využití, SW

dostupný přes web.
Předání výsledků v rámci průběžné zprávy projektu.

g) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2017-12-31

h) Číslo, název a cíl etapy

02 - Adaptace modelů a softwarových nástrojů
ZČU: Adaptace akustických a jazykových modelů, vývoj softwaru pro úpravy výstupu z OCR metodami zpracování přirozeného jazyka, testy metod detekce tématu
UK: Adaptace repozitáře, vývoj metod hloubkové analýzy textu pro potřeby archivu
USTR: Anotace nahrávek, konzultace metadatové struktury, příprava dokumentů ke zpracování

i) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2017-01-01

j) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2017-12-31

k) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

l) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

D – článek ve sborníku (plán ZČU: konference EU/ konference TSD 2017)
D – článek ve sborníku (plán ZČU: konference EU/ konference TSD 2017)
D – článek ve sborníku (plán UK: EU konference 2017)

m) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN.
Předání výsledků v rámci průběžné zprávy projektu.

n) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2017-12-31

o) Číslo, název a cíl etapy

03 Integrace modulů a softwarových nástrojů

ZČU: Vývoj rozhraní mezi vyhledávacím modulem a repozitářovým softwarem, integrace metod zpracování přirozeného jazyka, OCR a detekce tématu do výsledného softwaru HIDOAR

UK: Integrace systému HIDOAR, testy metod hloubkové analýzy textu pro potřeby archivu

USTR: Příprava nahrávek a dokumentů z jiných zdrojů, předběžné užitvatelské testování jednotlivých softwarových modulů

p) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

2018-01-01

q) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

2018-12-31

r) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

průmyslový výzkum

s) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

R – sw HIDOAR

D – článek ve sborníku (plán ZČU: konference INTERSPEECH/SPECOM/ICASSP 2018)

D – článek ve sborníku (plán UK: EU konference 2018)

t) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN.

Popisu funkčnosti výsledku typu R společně s licenčními podmínkami pro využití, SW dostupný přes web.

Předání výsledků v rámci průběžné zprávy projektu.

u) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

2018-12-31

v) Číslo, název a cíl etapy

| |
|---|
| 04 Testování a ladění integrovaného archivu, rozšíření dat a metadatového popisu<br>ZČU: Testy integrovaného systému HIDOAR, testy a úpravy metod rozpoznávání, indexace, vyhledávání, zpracování přirozeného jazyka, OCR a detekce tématu<br>UK: Závěrečné ladění a úpravy integrovaného systému HIDOAR, testy a úpravy metod hloubkové analýzy textu<br>USTR: Uživatelské testování integrovaného systému, metadatový popis dat s využitím softwaru HIDOAR |

w) Datum zahájení řešení etapy (*ve formátu: RRRR-MM-DD*)

| |
|---|
| 2019-01-01 |

x) Datum ukončení řešení etapy (*ve formátu: RRRR-MM-DD*)

| |
|---|
| 2019-12-31 |

y) Převažující typ výzkumu (*základní výzkum, průmyslový výzkum, vývoj*) při řešení etapy

| |
|---|
| průmyslový výzkum |

z) Výsledky etapy (*součet výsledků za všechny etapy musí odpovídat výčtu všech očekávaných výsledků projektu podle bodu č. 6 Popisu projektu*)

| |
|---|
| D – článek ve sborníku (plán ZČU: konference EU/ konference TSD 2019)<br>D – článek ve sborníku (plán ZČU: konference INTERSPEECH/SPECOM/ICASSP 2019)<br>D – článek ve sborníku (plán UK: konference USA/Kanada 2019) |

aa) Forma zpracování a předání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace*)

| |
|---|
| Vloženo do RIV, publikace evidována v příslušné databázi pod ISBN nebo ISSN.<br>Popisu funkčnosti výsledku typu R společně s licenčními podmínkami pro využití, SW dostupný přes web.<br>Předání výsledků v rámci průběžné zprávy projektu. |

bb) Termín odevzdání výsledků etapy (*v souladu s podmínkami pro předávání výsledků, uvedenými v příloze č. 7 zadávací dokumentace; ve formátu: RRRR-MM-DD*)

| |
|---|
| 2019-12-31 |

**10. Uvedení oponentů projektu, se kterými uchazeč nesouhlasí z důvodů možné podjatosti při hodnocení předloženého projektu** *(lze uvést max. 3 osoby nebo pracoviště)*.

| |
|---|
| prof. Ing. Jan Nouza, CSc., Ústav informačních technologií a elektroniky, Technická univerzita v Liberci |