MCLAAS INTEGROVANÝ SYSTÉM VYHLEDÁVÁNÍ VE VÍCEJAZYČNÉM AUDIOARCHÍVU uživatelská a instalační příručka



Obsah

1	Úvod	2
2	Certifikát a jeho instalace	3
	2.1 Instalace certifikátu v prohlížeči Internet Explorer	3
	2.2 Instalace certifikátu v prohlížeči Mozilla Firefox	5
	2.3 Instalace certifikátu v prohlížeči Google Chrome	8
3	Základní ovládání	13
4	Ovládání přehrávače	14
5	Instalace	16



1 Úvod

Tento software slouží pro vícejazyčné (české a anglické) vyhledávání relevantních slov či krátkých frází v archivu přeživších Holocaustu, spravovaném USC (University of Southern California) Shoah Foundation Institute¹. Tento archiv obsahuje více než 110 tisíc hodin záznamů v 32 jazycích, přičemž přibližně polovina těchto rozhovorů je vedena v angličtině. Česká část archivu obnáší zhruba jeden tisíc hodin.

Pro účely vyhledávání v softwaru systému MCLAAS jsou česká a anglická řečová data nejprve zpracována příslušným softwarem rozpoznávání řeči (SEASR-CZE - viz², resp. SEASR-ENG - viz³) a poté je vytvořen tzv. index, což je strojová reprezentace rozpoznaných promluv, která umožňuje co nejrychlejší vyhledání požadovaného slova či fráze. Oba vyhledávací systémy pracují s indexem založeným na slovní i fonémové reprezentaci, což umožňuje, kromě jiného, vyhledávat i slova mimo slovník.

Pro křížové vyhledávání (dotaz v češtině, data/rozhovory v angličtině a češtině) v softwaru byl použit systém překladu dotazu. Implementace byla provedena jako zvláštní verze systému MTMonkey⁴. Dotaz je klientem (tj. vyhledávačem, který jej získá od uživatele v poli pro zadání dotazu) poslán jako JSON-RPC požadavek aplikačnímu serveru, který je následně zašle příslušnému "workeru", v tomto případě překladovému systému pro překlad krátkých dotazů z češtiny do angličtiny. Ve výsledném formátu, ve kterém je překlad zaslán zpět klientovi, je opět ve formátu JSON obsažen i překlad dotazu, který je následně zpracován anglickou verzí vyhledávače, a výsledky jsou pak zobrazeny uživateli (spolu s výsledky z české části archívu).

Aplikace je dostupná na adrese https://amalach.zcu.cz. Nezabezpečená verze aplikace (protokol HTTP) není na serveru k dispozici a pokus o připojení k ní skončí sdělením internetového prohlížeče, že takovou stránku nenalezl.

¹http://dornsife.usc.edu/vhi/

²http://www.kky.zcu.cz/cs/sw/SEASR-CZE

³http://www.kky.zcu.cz/cs/sw/SEASR-ENG

⁴http://ufal.mff.cuni.cz/mtmonkey

2 Certifikát a jeho instalace

Pro používání aplikace je nutné, aby byl ve vašem prohlížeči k dispozici váš certifikát. O jeho specifikaci vás aplikace sama požádá.

Pokud potřebný certifikát ve výběru chybí, budete ho muset nejprve do svého prohlížeče jednorázově nainstalovat a pak již přihlášení probíhá automaticky bez nutnosti zadávat jméno a heslo. Bez úspěšné instalace klientského certifikátu se vůbec nepodaří k aplikaci připojit. Po instalaci zřejmě bude nutné ukončit a opět otevřít internetový prohlížeč.

Pokud zvolíte nesprávný certifikát, prohlížeč ohlásí chybu připojení SSL.

2.1 Instalace certifikátu v prohlížeči Internet Explorer

Naimportujte Váš klientský certifikát. Spusťte client_vašejméno.p12.

1.7.8 C	Q 3 -						x
Coo V l + cert				Prohledat: cert	-	-	Q
Uspořádat 👻 🗂 Nainstalovat PFX	✓ Sdílet s ▼	Vypálit	Nová složka			•	0
🔶 Oblíbené položky	Název položky	^		Datum změny	Тур	Velikost	
🖳 Naposledy navštívené	🔄 ca.crt			26.3.2015 13:29	Certifikát zabezpe	3 kB	
📃 Plocha	🦻 client_valentat.	512		31.3.2015 17:23	Personal Informati	5 kB	
🚺 Stažené soubory							
😌 Dropbox							
🕌 Disk Google 📃							
C Knihovny							
Dokumenty							
J Hudba							
Videa							
🍓 Domácí skupina							
1 Počítač							
🏭 Místní disk (C:)							
👝 Data (D:)							
👝 KINGSTON (G:)							
🚽 zcu.cz#user.valentat (\\AFS) (
🖙 public (\\konos-pole2-kkv.fa 🍸							
client_valentat.p12	Datum změny: 31.	3.2015 17:23	Date	ım vytvoření: 31.3.2015	17:36		
Personal Information Exchar	nge Velikost: 4,0	5 kB					

Obrázek 1: Internet Explorer: Výběr souboru s certifikátem

Tlačítkem další se doklikejte až k zadání hesla a zadejte heslo, které jste k certifikátu obdrželi.



Z důvodu zajištění zabezpečení byl privá	itní klíč chráněn heslem.
Zadejte heslo k privátnímu klíči.	
Heslo:	
pri pouzia privatnino kilc aplikad Qznači tento kilč jako exportoval přenést Zghrnout všechny rozšířené vlast	vyzvani k potvrzeni akce) telný; později bude možné klíč zálohovat či nosti
Další informace o <u>ochraně privátních klíčů</u>	

Obrázek 2: Internet Explorer: Zadání hesla k certifikátu

Tlačítky Další/Dokončit import dokončete.

Úložiště o	ertifikátů
Úložiš	tě certifikátů jsou oblasti systému, kde jsou uloženy certifikáty.
Systé umíst	m Windows může automaticky vybrat úložiště certifikátů, nebo můžete zadat Ení certifikátu.
۲	Automaticky vybrat úložiště certifikátů na základě typu certifikátu
C	Všechny certifikáty umístit v následujícím úložišti
	Úložiště certifikátů:
	Procházet
Další infori	nace o <u>úložištích certifikátů</u>
	< Zpět Další > Stor

Obrázek 3: Internet Explorer: Dokončení importu certifikátu

Do prohlížeče zadejte adresu https://amalach.zcu.cz a budete vyzvání k výběru Vašeho certifikátu. Po potvrzení se zobrazí stránka AMALACH.





Obrázek 4: Internet Explorer: Potvrzení použití certifikátu

2.2 Instalace certifikátu v prohlížeči Mozilla Firefox

Otevřete možnosti aplikace





Obrázek 5: Mozilla Firefox: Možnosti aplikace

Menu Rozšířené, záložka Certifikáty, stiskněte tlačítko Certifikáty.



Obrázek 6: Mozilla Firefox: Možnosti aplikace (záložka Rozšířené)

Na záložce Osobní stiskněte Importovat.





Obrázek 7: Mozilla Firefox: Certifikáty

Vyberte soubor klientského certifikátu a stiskněte Otevřít.

😻 Soubor s certifikátem k imp	portu		×
💮 💮 🕨 🕶 cert	▼ 	Prohledat: cert	2
Uspořádat 🔻 Nová složka	3	8==	• 🔳 🔞
😠 Public (\\NAS) (r 📥	Název položky 🔺	Datum změny	Тур
valentat (\\kono: ♀ C na TOM-NB ♀ D na TOM-NB ♀ E na TOM-NB ♀ F na TOM-NB ♀ N na TOM-NB	👼 client_valentat.p12	31.3.2015 17:23	Personal Info
📭 Síť			
🔎 cert	•		Þ
<u> </u>	, souboru: client_valentat.p12 ▼	PKCS12 Soubory (*.p1	2;*.pfx) Storno

Obrázek 8: Mozilla Firefox: Výběr souboru s certifikátem

Zadejte heslo.



Dialog zad	lání hesla				×
Prosím certifiká	zadejte heslo, kte tu.	ré bylo pou	žito pro za	išifrování zi	ilohy
Heslo:	•••••				
			OK	Zruš	it

Obrázek 9: Mozilla Firefox: Zadání hesla k certifikátu

Zadejte adresu https://amalach.zcu.cz a budete vyzvání k výběru Vašeho certifikátu. Po potvrzení se aplikace zobrazí.



Obrázek 10: Mozilla Firefox: Potvrzení certifikátu

2.3 Instalace certifikátu v prohlížeči Google Chrome

Pro instalaci certifikátu v internetovém prohlížeči Google Chrome přejdeme nejprve do nastavení prohlížeče. To lze např. volbou "Nastavení" v menu prohlížeče v jeho pravém horním rohu.



→ C ♠	Chome://settings		
/ U II			
hrome	Nastavení	Prohledat nastavení	
istorie	Přihlaste se		
zšíření	Jste přihlášeni jako petr.salajka@gmail.com. Synchronizaci dat můžete spravova	t pomocí <u>Hlavního panelu</u>	
istavení	Google. Odpojit účet Google Rozšířená nastavení synchronizace		
ormace	Při spuštění		
	Otevřít stránku Nová karta		
	Pokračovat tam, kde jste přestali		
	Otevřít konkrétní stránku nebo skupinu stránek. Skupina stránek		
	Vzhled		
	Získat motivy Obnovit výchozí motiv		
	Zobrazit tlačítko Domovská stránka		
	facebook.com/login.php;gmail.com Změnit		
	Vždy zobrazovat lištu záložek		
	Vyhledávání		
	Nastavit vyhledávač, který se používá při vyhledávání v omniboxu.		
	Google Spravovat vyhledávače…		
	Lidé		
	💭 Petr (aktuální)		
	the second s		

Obrázek 11: Google Chrome: Nastavení

Následně zadáme do vyhledávacího pole výraz "certifikáty" nebo "SSL". Objeví se tlačítko "Spravovat certifikáty...", na které klikneme. Otevře se okno "Certifikáty" a pokud ještě žádné certifikáty importované nemáte, bude vypadat stejně jako na následujícím obrázku. Pak je třeba zvolit "Importovat...". Tím se spustí průvodce importem certifikátu.

nýšlen	ý účel:	<vše< th=""><th>></th><th></th><th></th><th></th></vše<>	>			
sobní	Ostatní uživa	itelé	Zprostředkující certifika	ční autority	Dův	ěryhodné kořenové ce 🕚
Vysta	iveno pro		Vystavitel	Datum		Popisný název
porto	vat) Expor	tovat	Odgbrat			Upřesnit
porto amýšle	vat) Expor	tovat fikátu	Odgbrat			Upřesnit
porto amýšle	rat) Expor	tovat fikátu	Odgbrat			<u>Upřesnit</u> Zobrazit

Obrázek 12: Google Chrome: Certifikáty

První okno průvodce je pouze informační a stačí stisknout tlačítko "Další >". Ve druhém okně pomocí tlačítka "Procházet..." vyberme soubor s naším



certifikátem. Dialogové okno výběru souboru nemusí nejprve žádné soubory ukázat. Ověřte prosím, že filtr zobrazovaných souborů odpovídá typu souboru, ve kterém je váš certifikát.

Zadejte soubor, který drazte importovat. <u>B</u> ištev souboru: E:VProjektr/WAKLACH/certs/clent_sals/ka.p12 Porsárkia: Y Jednom souboru je možné uložit více než jeden certifiká formátně: Pormát Personal Informaton Exchange - PKCS č. 12 (PFK, P12) Certifikáty standardu Crystographic Message Syntax Standard - Serialzované úložiště certifikátů (SST)	
Nárev soubonu: E: VProjektry UAALACH/certs/clent_salspla.p12 Popularika: V Jachon souboru je možiné uložit více než jeden certifiká formáteči: Formát Personal Information Exchange - PKCS č. 12 (PFX, P12) Certifikáty stardadu Cryptographic Message Syntax Standard - Serialzované úložiště certifikátů (SST)	
E: (Projekty UMALACH (perts)diemt_salaşka, p12 Poznămika: V şisdrom souboru je možné uložit více než jeden certifiká formátek: Formát Personal Information: Euchange - PKCS č. 12 (PEV, P12) Certifikáty standardu Cryptographic Message Syntax Standard - Serializované úložiště certifikátů (SST)	
Poznárska V jednom souboru je možné ubžit více než jeden certifiká formáteci: Pomit Personal Information Exchange - PKCS č. 12 (PFK, P12) Certifikáty standardu C rystographic Message Syntax Standard - Senalzované úložště certifikátů (SST)	Procházet.
Formát Personal Information Exchange - PKGS č. 12 (PFX, P12) Certifikáty standardu Gryptographic Message Syntax Standard - Serializované úložiště certifikátů (SST)	t v následujícíc
Certifikáty standardu. Oryptographic Message Syntax Standard - Senalizované úložiště certifikátů (SST)	
Serializované úložšítě certifikátů (SST)	чксs č. 7 (Р7В
n Michael and a factor is a second state	
o mine and the base of the	
Daisi informace o formatech souboru certifikatu	

Obrázek 13: Google Chrome: Výběr souboru s certifikátem

Následně zadejte heslo.

ůvodce	importem certifikátu
Heslo	
z	důvodu zajištění zabezpečení byl privátní klíč chráněn heslem.
z	adejte heslo k privátnímu klíči.
	Heslo:
	••••••
	Qynačit tento klič jako exportovatelný; pozděj bude možné klič zálohovat č přeněst Zghrnout všechny rozšířené vlastnosti
Další ir	iformace o <u>ochraně privátních klíčů</u>
	< Zpět Další > Storno

Obrázek 14: Google Chrome: Zadání hesla k certifikátu

V dalším kroku zvolte úložiště certifikátu. Pokud si nejste jistí, výchozí volba může být nejvhodnější.



i jsou uloženy certifik	áty.
iložště certifikátů, n	ebo můžete zadat
ú na základě typu cer	rtifikátu
cím úložišti	
	Procházet
	Jozste certitikato, n I na základě typu ce cim úložští

Obrázek 15: Google Chrome: Výběr úložiště certifikátu

Nyní už vás čeká pouze shrnutí celého importu a pokud jste s vaším nastavením spokojení, můžete import dokončit.

Dokončení Průvodc certifikátu	e importem
Certifikát bude naimportován p Dokončit.	o kliknutí na tlačítko
Zadali iste následující nastaven	é:
Úložiště certifikátů vybrané u Obsah	ävatelem Osobní PEV
Název souboru	E:\Projekty\AMAI
<	,

Obrázek 16: Google Chrome: Shrnutí importu certifikátu

Mělo by se objevit okno informující vás o úspěšném importu certifikátu.

Průvodce ir	mportem certifikátu 🛛 🔀
i	Import proběhl úspěšně.
	ОК

Obrázek 17: Google Chrome: Okno potvrzující úspěšný import certifikátu

Pokud import proběhl úspěšně, certifikát již bude ve výčtu uveden a vy můžete začít aplikaci používat.



obní Ostatní uživa	atelé Zprostředkující certifik	Zprostředkující certifikační autority Důvěryhodné					
Vystaveno pro	Vystavitel	Datum		Popisný název			
alsalajka	amalach.zcu.cz	7.4.201	16	<Žádný>			
ortovat) Expon	tovat) Odgbrat			Upřes			

Obrázek 18: Google Chrome: Certifikáty (certifikát je již k dipozici)



3 Základní ovládání



Do kolonky "Zadejte dotaz" napište slova, která se mají ve vámi hledané nahrávce vyskytovat. Následně stiskněte klávesu "ENTER" nebo myší ikonu lupy napravo od dotazu. Systém bude automaticky hledat i skloňované tvary vámi zadaných slov. Pokud si automatické skloňování nepřejete (chcete hledat pouze vámi zadaný tvar), uzavřete slovo do uvozovek. Slova, která začínají znakem plus se ve výsledku mohou, ale nemusí vyskytovat. Systém vyhledává nad automatickým přepisem nahrávek. Velikost písmen tedy nehraje roli.

Pokud si přejete vyhledat slovo v anglickém archivu, ale nevíte, jak se váš dotaz anglicky napíše, napište ho česky a použijte pro překlad tlačítko napravo od "lupy". Ikona dále napravo ukazuje směr překladu. <u>"CS</u> \rightarrow EN" tedy znamená, že se bude překládat z češtiny do angličtiny. Červené podtržení značí jazyk archivu, který se bude prohledávat. Všimněte si, že když provedete překlad vašeho dotazu z češtiny do angličtiny, přesune se automaticky podtržení pod "EN". Do angličtiny přeložený dotaz se tedy bude vyhledávat v anglickém archivu. Pokud si přesto přejete hledat přeložený dotaz v původním jazyce, můžete to kliknutím na tuto nabídku změnit.

Poslední tlačítko (první zprava) umožňuje vstup do nastavení systému. Výchozím jazykem ve kterém se vyhledávání provádí je jazyk prostředí. Jazyk prostředí lze změnit v nastavení aplikace.



4 Ovládání přehrávače



Pro ukázku si představme, že chceme vyhledat slovo "československo". Do kolonky "Zadejte dotaz" tedy napíšeme "československo" a stiskneme tlačítko "s lupou". Tím se přesuneme na obrazovku s výsledky vyhledávání.

AMALACH _{search}						
Výskytů: 500+	Seřadit podle: % 00:00					
ČESKOSLOVENSKA Ruzena Blechova (12. 11. 1996)						
ČESKOSLOVENSKU Ruzena Blechova (12. 11. 1996)	100 %					
ČESKOSLOVENSKU Eliska Fábryová (11. 11. 1996)	100 %					
ČESKOSLOVENSKA Anna Mrazkova (3. 9. 1996)	100 %					
ČESKOSLOVENSKA Herta Kohnova (4. 3. 1996)	100 %					

V levém sloupci máme seznam výsledků spolu se jménem řečníka, datem nahrávání a číselným ohodnocením, jak moc si je systém jistý svým rozhodnutím, že výsledek odpovídá vašemu dotazu (od 0 do 100%).



československo	۹	ネ	\underline{CS} → EN	
Ruzena Blechova (European Holocaust, 1939-1945)				

Napravo nahoře je panel základního ovládání (viz základní ovládání) a pod ním jméno řečníka spolu s pojmenováním archivu.



Uprostřed je samotný přehrávač videa. Tenká červená linka pod obrazem ukazuje pozici v nahrávce. Zelené značky ve stejné oblasti pak ukazují výsledky hledání v rámci dané nahrávky. Zvolíme je kliknutím myši. V nahrávce se lze také volně pohybovat kliknutím na libovolné umístění. Hlasitost lze měnit pomocí opakovaného stisku klávesových kombinací $CTRL+\uparrow$ a $CTRL+\downarrow$.

Výchozím nastavením je, že se výsledek přehraje vždy s určitým okolím (před a po místě, kde byl nalezen). Pokud chceme jednorázově přehrát výsledek s větším okolím, použijeme tlačítko s ikonou " $<\cdots>$ ". Po každém stisknutí se přehrávané okolí výsledku zvětší o 5 sekund na obě strany. Jednorázové zvětšení platí pouze pro právě přehrávaný výsledek. Poslední tlačítko (první zprava) umožňuje opakované přehrávání. Prvním kliknutím nastavíme opakované přehrávání seznamu výsledků. Druhým kliknutím pak opakované přehrávání daného výsledku. Třetím kliknutím opakování opět zrušíme.

Napravo od přehrávače je potom výčet některých vlastností dané nahrávky jako např. jméno řečníka, jméno reportéra nebo jazyk nahrávky.



5 Instalace

Pro instalaci se předpokládá počítač s linuxovou distribucí. Doporučuje se aktuální Debian Stable. Dále je třeba nainstalovat aplikaci Docker podle návodu na oficiálních stránkách projektu⁵. Postup sestavení pracovní databáze je popsán v dokumentaci softwarů SEASR-CZE a SEASR-ENG. Předpokládejme tedy, že máme k dispozici funkční databázi MongoDB a archivovaný obraz (Docker Image) serverové aplikace složené ze softwarů MCLASS a WFBAS - amalach_prototype.tar.gz. MCLASS nedokáže sám pracovní databázi prohledávat a WFBAS na druhé straně nepředpokládá pokročilé uživatelské rozhraní. Proto jsou pro naše potřeby oba softwary distribuovány pohromadě.

Nejprve je třeba v archivu upravit soubor opt/pixla/amalach.json, kde je popsán způsob přístupu k databázi. Především db_uri:

"db_uri": "mongodb://hydra3.kky.zcu.cz/amalach3/px01",

Dále je třeba přidat do archivu soubory:

- TLS certifikát opt/pixla/ca/amalach.zcu.cz.pem
- TLS klíč opt/pixla/ca/serverkey.dec.pem
- TLS key-chain opt/pixla/ca/tcs-ca-bundle.pem (musí obsahovat CA, která vydává klientské certifikáty)

Software TMODS:ENG-CZE, který umožňuje překlad uživatelských dotazů mezi angličtinou a češtinou běží na vzdáleném počítači. Pro správnou spolupráci je třeba nastavit uri v souboru opt/pixla/pixla_server.py:

```
(r"/translate", ProxyHandler, {
    'uri': 'http://lindat.mff.cuni.cz/services/moses/request',
    'concat': False
}),
```

Dále popsané příkazy vkládáme do příkazové řádky. Nejprve importujeme archiv a vytvoříme tak Docker image:

docker import amalach_prototype.tar.gz amalach/prototype

Vytvoří se Docker image s názvem amalach/prototype. Následně lze již aplikaci spustit příkazem:

⁵http://www.docker.com/

docker run -i -t --net=host --entrypoint /usr/bin/supervisord \ --workdir=/opt/pixla amalach/prototype -nc /opt/pixla/supervisord.conf

Význam všech použitých přepínačů je součástí dokumentace Dockeru. Význam přepínačů /usr/bin/supervisord snadno zjistíme, pokud místo příkazu výše zadáme např.

docker run -i -t --net=host --entrypoint /usr/bin/supervisord \setminus --workdir=/opt/pixla amalach/prototype --help

