Český rozhlas

# Protokol z předběžné tržní konzultace

## v souladu s § 33 Zák. č. 134/2016 Sb., o zadávání veřejných zakázek

## „Indexace obsahu a vyhledávání, a další služby spojené s automatizací textu “

**Datum a čas konání:** středa 23. ledna 2019 od 15:00 hodin

**Místo konání:** místnost č. C 425, Vinohradská 12,120 00 Praha 2

**Přítomní zástupci za zadavatele (ČRo):**

- Jiří Špaček
- Jindřich Kubelík
- Daniel Felix Hrouzek
- Martin Zadražil
- Eva Gottová

**Přítomní zástupci za dodavatele:**

- **Geneea Analytics s.r.o.**
- Petr Hamerník
- Jiří Hana
- **INOVATIKA s.r.o.**
- Tomáš Dočkal
- **TOVEK, spol. s r.o.**
- Miroslav Nečas
- Miroslav Wiedermann
- **NEWTON Technologies, a. s.**
- Petr Herian
- Ludmila Tydlitátová

**Zápis z průběhu jednání:**

1) Úvod

- Úvodem podepsali všichni přítomní zástupci zadavatele i dodavatelů Prezenční listinu účastníků, která tvoří přílohu č. 1 tohoto písemného protokolu, čímž zároveň vyjádřili souhlas se skutečností, že z této předběžné tržní konzultace bude pořízen zvukový záznam, který bude sloužit výhradně k účelu vypracování písemného protokolu.
- Eva Gottová dále informovala všechny zúčastněné osoby, že písemný protokol bude následně všem účastníkům předběžné tržní konzultace zaslán ke schválení a poté bude zveřejněn na profilu zadavatele. Dále upozornila zástupce dodavatele, že má právo označit informace, které by případně mohly v budoucnu narušit hospodářskou soutěž, za důvěrné. Takové informace následně nebudou součástí veřejného písemného protokolu.

2) Průběh jednání

- Jiří Špaček – Prezentace projektu „můjROZHLAS“ jako jednoho ze tří pilířů internetu ČRo (rozhlas.cz, irozhlas.cz, můjrozhlas.cz) pro kterého hledáme dodavatele, který nám přinese řešení „chytrého“ vyhledávání.
- Samotné vyhledávání je rozděleno do dvou částí na zpracování archivu kontinuálního vysílání a zpracování živého vysílání:
- Archiv kontinuálního vysílání stanic má ČRo od roku 2003, a je nutné jej zpracovat a provést jeho transkripci (netýká se této veřejné zakázky). Jedná se o transkripci včetně diarizace a identifikace řečníků, následné rozřazení dle konkrétních stanic, jejich pořadů a epizod.

1.7.2019 13:50:32
CN: Eva Gottová
O: Český rozhlas
SN: 0xB0002F
S časovým razítkem

- Tyto texty v určité kvalitě přepsání bude nutné indexovat (z části na straně ČRo) a v těchto datech bezpečně vyhledávat dle řečníků (př. prezidenta ČR Miloše Zemana hovořícího na téma „Temelín“), tyto obsahy štítkovat dle klíčových slov a tematizovat.
- To samé chce zadavatel provést se živým vysíláním online.
- Do budoucna ČRo uvažuje (taktéž nebude součástí této veřejné zakázky) u některých epizod provádět automatické shrnutí v podobě plynulého textu pro „promo“ těchto obsahů, ale i pro budoucí službu „text to speech“.
- Do aktuálně připravované veřejné zakázky ovšem spadá výhradně vyhledávání, štítkování a tematizaci, resp. zařazování epizod pod určitá témata.
- Daniel Hrouzek – Prezentace technické stránky zpracování
- Předány informace o hlavním úložišti dat textového charakteru (elastic search), tedy určitý „index“ bude již existovat, pravděpodobně nepůjde o index, nad který bude pracovat vybraný dodavatel.
- Přístup k původním datům v elastic search bude jak přímý, tak prostřednictvím určitého „obalujícího“ API, nebo bude možné vytvořit přístup ad hoc na základě požadavku vybraného dodavatele.
- Transkripce bude probíhat nad samotným archivem, ale taktéž nad on-line vysíláním. To znamená, že i tento obsah bude následně nutné indexovat, zařazovat pod témata a štítkovat.
- Každá část těchto dat bude taktéž opatřena časovou značkou, časová osa je to hlavní, s čím tento projekt pracuje.

TOVEK, spol. s r.o.

- Upozornění na význam úrovně kvality dat přepisu a vliv chybovosti textu. Půjde o plynulý stream slov nebo jednotlivá slova a věty? Budou tyto věty ukončeny? Důležitost přesnosti přepisu v procentech (70 – 90%)?

ČRo

- Vybraný dodavatel obdrží transkripci v takové míře, v jaké ji ČRo bude mít k dispozici. Nyní není možné definovat přesnost přepisů v procentech. ČRo v současné době připravuje veřejnou zakázku „Analytika mluveného slova„, jejímž předmětem bude právě transkripce, diarizace a identifikace mluvčích. Součástí zadávací dokumentace budou podmínky pro budoucího dodavatele a limity, ke kterým se chce ČRo dostat, a to včetně přesnosti přepisů na 95%. Co se týká indexace a vyhledávání, je možné, že s vylepšeným algoritmem bude nutné data „přeindexovat“ v čase.

INOVATIKA s.r.o.

- Dotaz na vnější lidské korektury?

ČRo

- Tyto lidské korektury budou u vybraných pořadů. Samotný přepis bude mít pravděpodobně tři základní úrovně (on-line přepis ze živého vysílání – méně přesný/hloubkový přepis – transkripce po skončení pořadu/ruční zásah editora). Zásah editora má dva stupně, a to základní korekturu, nebo kompletní úpravu. Vybraný dodavatel bude mít vždy označeno, o jakou úroveň se jedná, nejčastěji se však bude jednat o on-line přepisy.

INOVATIKA s.r.o.

- Data, které vybraný dodavatel obdrží, může považovat za finální a pracovat s nimi? Nebo bude zadavatel po dodavateli požadovat ještě jejich úpravu, např. odstranění neidentifikovatelných/podezřelých slov?

ČRo

- To záleží na technických návrzích řešení, zadavateli jde o co nejlepší výsledek.

Geneea Analytics s.r.o.

- Dotaz na způsob měření přesnosti 95 %? Počet správných slov….

ČRo

- Vychází se z vícera parametrů, správný počet slov, počet správně přepsaných slov, správné koncovky, interpunkce, diakritika… Snahou zadavatele je samozřejmě vysoutěžit co nejvyšší úroveň.

TOVEK, spol. s r.o.

- Zdůraznění významu kvality přepisů pro činnost indexace, tedy do jaké míry se bude moci vybraný dodavatel na kvalitu dat spolehnout.

ČRo

- V případě dvoufázových přepisů to samozřejmě i pro dodavatele indexace bude znamenat dva kroky.

Geneea Analytics s.r.o.
- V takové situaci to znamená použití dvou různých systémů, což bude mít na samotnou nabídku rozhodující dopad. Budou k dispozici informace o vysoutěženém systému na zmiňovanou analytiku mluveného slova?

ČRo
- Tato zakázka na „vyhledávání“ bude probíhat cca do dvou měsíců a během té doby bude již vyhlášena zakázka na „Analytiku mluveného slova“. Pokud bude zakázka na „Analytiku“ ukončena dříve, pak zadavatel může poskytnout konkrétní informace, pokud nikoli tak budou k dispozici pouze informace obecné.

INOVATIKA s.r.o.
- Dotaz na první fázi zpracovaných dat z on-line přepisu, co s nimi bude dále dělat, když se naindexují.

ČRo
- Uvedení příkladu pořadu Lucie Výborné (9:00 – 10:00) rozhovor s „nějakým horolezcem“, tak po zadání klíčového slova „horolezec“ a času 9:15 hod, už by měl být tento pořad zařazen s přepisem prvních 15 minut.

TOVEK, spol. s r.o.
- Dotaz, zda zadavatel si bude štítky vybírat a následně editovat jejich seznam sám, neboť může jít o stovky až tisíce? Témata budou spíše obecnější, má zadavatel již nějakou představu?

ČRo
- Ano, zadavatel si bude štítky vybírat sám a předpokládá jejich množství spíše v počtu tisíců. Taktéž zadavatel si bude na počátku téma definovat, ale samozřejmě témata jsou proměnná a neustále vznikají nová.

NEWTON Technologies, a. s.
- Dotaz na rozdíl mezi štítkováním a tematizováním?

ČRo
- Např. do tématu „Dálnice D1“ zařadíte i stížnosti okolních obcí. Štítek je, zda se v daném textu slovo objevilo a má tam svou váhu. Téma je širším pojmem, např. téma „17. listopad“ může zahrnovat i divadelní hry nebo projev prezidenta na Václavském náměstí.
- V rámci webové aplikace bude zadavatel sám určovat jednotlivá aktuální témata a sám bude definovat, která témata budou její součástí, tak aby každé téma nebylo součástí aplikace automaticky.

INOVATIKA s.r.o.
- Správa těchto témat bude součástí této veřejné zakázky nebo si je bude spravovat sám zadavatel?

ČRo
- Ideálně zadavatel nadefinuje nějaká témata a dodaný systém by měl být schopný dodat témata nová viz příklad facebookových clusterů a s tím spojené vytvoření nového segmentu uživatelů…

TOVEK, spol. s r.o.
- Skutečnost, že je nějaké téma zajímavé musí určit člověk…

ČRo
- Zadavatel počítá s pracovní pozicí hlavního editora tohoto projektu, který bude mít na starosti stanovení témat i klíčových slov.
- Seznamy štítků včetně jednotlivých témat

Daniel Hrouzek
- Seznamy štítků a témat dá zadavatel k dispozici vybranému dodavateli via API. Zároveň zadavatel bude očekávat od dodavatele návrhy nových témat, nových klíčových slov…

Geneea Analytics s.r.o.
- Dotaz na klíčová slova typu „Miloš Zeman“, v takém případě nepůjde o tisíce, ale desetitisíce. Otázka standardizace viz „DPH“ vs. „Daň z přidané hodnoty“…

ČRo

- Ano, pravděpodobně půjde o desetitisíce, ideálně pro zadavatele do 20.000 klíčových slov, jde o vysílání za období let 2003 - 2018.
- Do jisté míry záleží na dodavateli, který bude v rámci veřejné zakázky „Analytika mluveného slova“ vysoutěžen.

Geneea Analytics s.r.o.
- Budou k dispozici metadata zadavatele?

Daniel Hrouzek
- Ano, dodavatel bude mít všechna data s popisy, významem a strukturou k dispozici a bude moci je používat stejným způsobem jako zadavatel.

Geneea Analytics s.r.o.
- Otázka na samotné využití indexace, web a redakční práce?
- Prostředí je jednotné, tedy přístup je jednotný zvenku i zevnitř?

ČRo
- Dodavatel bude přepisovat vše, ale jen určitá část bude zveřejněna, neboť toto závisí na určitých právech. Některé pořady podléhají zákazu zveřejnění na internetu a zákazu přepisů obecně
- Ano, přístup do prostředí je jednotný.

Geneea Analytics s.r.o.
- Bude mít zadavatel nějaké zvláštní požadavky např. ve vztahu k GDPR ?

ČRo
- Zadavatel by měl být schopen dohledat identitu jedince. Pokud někdo bude vyžadovat své vymazání z konkrétního pořadu nebo epizody, toho by měl být zadavatel schopen. Dodavatel by měl poskytnout technický prostředek, aby takové vyhledání bylo možné.

INOVATIKA s.r.o.
- Prezentace možnosti detekce míst, lidí a lokalit na základě klíčových slov. Detekce entit (lidé /místa) může být pro zadavatele zajímavá, neboť jistě zadavatel disponuje, byť třeba v omezené míře, nějakými průvodkami u jednotlivých audiozáznamů (minimálně u rozhovorů). Otázka, zda zadavatel již disponuje nějakými slovníky s mluvčími, otázka párování na konkrétní redaktory. Otázka schvalování nových jmen/míst…

ČRo
- Identifikace mluvčích by měla proběhnout v rámci zakázky „Analytika mluveného slova“. Identifikace může probíhat různými způsoby, může vycházet z databáze otisku hlasů, kterou bude zadavatel budovat. Začne se s lidmi v rámci ČRo, poté lidmi zvenčí, kteří hovoří nejvíce a postupně se budou zavádět jednotlivé profily dle předem stanoveného žebříčku. Identifikace mluvčích však nebude součástí veřejné zakázky na „Indexaci“.
- Zadavatel spíše nepočítá s lokací /identifikací míst.

Geneea Analytics s.r.o.
- Dotaz na štítkování viz příklad Pavel Kohout/Pavel Kohout/Pavel Kohout – bude zadavatel požadovat 3 štítky nebo jeden? Dtto „Jarda Jágr“, „Václav Klaus“; britská královna = Alžběta II.?

ČRo
- Musí být samozřejmě odpovídající počet štítků. Při zadání „britská královna“ do Googlu se pravděpodobně namixují výsledky dle očekávání toho, co uživatel hledá. Pokud toto uživatel zadá do systému dodavatele, jaké údaje uživateli vypadnou?

Geneea Analytics s.r.o.
- Pokud uživatel chce, aby vypadla Alžběta II, tak vypadne Alžběta II. Pokud uživatel zadá „britská královna“, jsou-li tam obě Elisabeth, vypadne Elisabeth I. a II. Jde o to, zda půjde o takový požadavek zadavatele, zda si zadavatel stanoví, že např. “Dálnice„ = „D1“ Ale to si zadavatel řeší sám ve svém systému. Bude zadavatel požadovat, aby štítky obsahovali i další/podrobnější údaje o „britské královně“ nebo jen definici osoby s určitým číslem, kterému zadavatel přiřadil název „Alžběta II./britská královna“?

ČRo
- Záleží na doporučení dodavatelů a koncepci navrhovaného řešení, přesto se zadavatel bude snažit nadefinovat své preference v ZD a cíle.

- Hodnotit se pravděpodobně bude cena a návrh konceptu řešení
- Spolupráce by měla být dlouhodobějšího charakteru, smlouvu zadavatel plánuje uzavřít na období 4 let.
- Obecné hledání bude vracet obecné výsledky a dále bude možnost parametrizovaného hledání.
- Zadavatel bude chtít zakazovat zobrazení některých výsledků a taktéž požadovat preferenci zobrazovaných údajů zadavatel, pokud to systém dodavatele bude umět.

ČRo

- Zadavatel vlastní pouze „index“, nikoli vyhledávací aplikaci
- U zadavatele je pouze „elastic search“ a součástí dodávky musí být „frontend/aplikace a to, co by se indexovalo.
- Zadavatel poskytne dodavateli přístup k „syrovým datům“ via úložiště, do kterého se nezapisuje (otázka dohody s dodavatelem)
- Otázka opravy a úpravy textu, které budou probíhat u nás, a dodavatel o tom bude informován.
- Ideální stav by byl systém, který bude schopen dělat opravy do češtiny, tedy co neopraví transkripce, opraví dodaný systém indexace.

TOVEK, spol. s r.o.

- Otázka smluvního vztahu. Není možné, aby zadavatel definoval odpovědnost dodavatele za garanci kvality vstupních údajů. Kvalita indexace bude výrazným způsobem ovlivněna vybraným dodavatelem z předchozí zakázky, tj. „Analytika mluveného slova. Otázka zdrojových kódů. Bude zadavatel požadovat řešení na míru? V takovém případě se zdrojové kódy předávají, v ostatních případech nikoli. Mělo by být následně zohledněno v zadávací dokumentaci, respektive ve smlouvě.

INOVATIKA s.r.o.

- Do jaké míry bude zadavatel požadovat analýzu textu a bude se pro zadavatele vytvářet speciální algoritmus na míru nebo do jaké míry chce zadavatel koupit nějaký balík slovníků.
- Rozdělení do dvou částí vyhledávání se standardními slovníky ČJ/štítkování se slovníky zadavatele, vyhledávání dle volného textu.

ČRo

- Toto je na zvážení do budoucna, prozatím není možné přesně odpovědět. Zadavatel se spíše kloní k již zavedeným slovníkům. Ještě před vyhlášením veřejné zakázky zadavatel zvažuje zveřejnit technickou specifikaci požadavků.

NEWTON Technologies, a. s.

- Otázka parametrů hodnocení kvality.

ČRo

- Zadavatel zvažuje nastavení hodnotících kritérií v určitém poměru cena/kvalita, prozatím bez bližší specifikace.

TOVEK, spol. s r.o.

- Otázka frontendu, včetně podpory pro daného editora.
- Otázka definice témat.

ČRo

- Interní zaměstnanci jistě nebudou pracovat v tom samém frontendu, jako externí uživatelé, ale obě skupiny budou využívat jednotné API
- Zadavatel hledá dlouhodobého partnera, se kterým bude pracovat na rozvoji daného systému.
- Zadavatel požaduje historii vyhledávání pro každého uživatele (do budoucna i hlasové vyhledávání).
- Zadavatel prozatím není připraven na vyhledávání dle zvuků, prozatím přichází v úvahu pouze práce s textem.

INOVATIKA s.r.o.

- Dotaz na objem dat.

ČRo

- Zadavatel disponuje přesným počtem hodin záznamů, není však rozděleno na hudba/mluvené slovo. Týká se 24 stanic a 24 hod. denně. Přesné počty budou známy až po realizaci zakázky na „Analýzu mluveného slova“, tyto údaje budou součástí zadávací dokumentace pro „Indexaci“.

- On-line transkripce bude zahájena ihned a samotné zpracování celého archivu je na cca 3 roky a rychlost zpracování samozřejmě závisí na vybraném dodavateli.
- Stávající archiv existuje v digitální podobě.
- Dodavatel nadefinuje hardware (dle požadovaného objemu dat zadavatele), který si zadavatel pořídí sám, hardware nebude součásti hodnocení veřejné zakázky.

Geneea Analytics s.r.o.
- Dotaz na možnost cloudové řešení pro samotné vyhledávání.

ČRo
- Zadavatel proti cloudovému řešení obecně nic nenamítá, ovšem jsou zde určitá rizika spojená se správou systému.
- Co se týká zpracování archivu, tedy posílání dat ke zpracování, jejichž objem bude značný, toto je čistě na dodavateli.
- Otázka kvality rychlosti přepisu

TOVEK, spol. s r.o.
- Důležitost počtu témat a jejich stanovení – musí být součástí zadávací dokumentace.
- Témata si musí zadavatel definovat sám, dodavatel dodá pouze technický nástroj.

ČRo
- Bude prací již zmiňovaného hlavního editora.

INOVATIKA s.r.o.
- Bude mít zadavatel nějaké specifické požadavky na vyhledávání? Fasety, možnost řazení dle typu dokumentů?

ČRo
- Zadavatel bude nad těmito možnostmi ještě interně jednat.
- K dotazům týkající se jazykových mutací zadavatel uvádí, že se bude týkat výhradně o češtinu, jiné jazyky nebudou součástí toto veřejné zakázky. Zároveň je nutné zdůraznit, že objem např. slovenštiny neustále narůstá.
- Co se týká transkripcí, tak požadavek na češtinu i slovenštinu bude součástí zadávacích podmínek.

Geneea Analytics s.r.o.
- Hledání češtiny může být i ve slovenštině. Český dotaz hledaný ve slovenštině…

NEWTON Technologies, a. s.
- Podoba transkripce anglického textu…Transkripce i přesto proběhne a systém se bude snažit přepsat cizojazyčný text vždy do „nějaké podoby“…

TOVEK, spol. s r.o.
- Otázka stanovení jednotlivých zón v rámci textu zadavatelem.

ČRo
- Podklady očekáváme od dodavatele.

NEWTON Technologies, a. s.
- Automatická detekce konce zpráv.

ČRo
- Systém by měl umět detekovat konce zpráv, neboť zadavatel těmito daty bude disponovat a předá je dodavateli indexace včetně časových razítek.

Geneea Analytics s.r.o.
- Jednotliví mluvčí budou odděleni?

ČRo
- Ano, identifikace mluvčích je součástí zadávacích podmínek pro „Analytiku slova“.

INOVATIKA s.r.o.
- Jaké bude následné využití ve vyhledávání?

ČRo
- Uveden příklad Miloše Zemana, bude možné vyhledat, kdy byl mluvčím a kdy bylo hovořeno o něm.

INOVATIKA s.r.o.
- Štítky tedy musí být rozděleny do dvou skupin a být schopny rozeznat, kdy hovořil mluvčí a kdy bylo hovořeno o něm…

ČRo
- Je otázkou, zda má smysl štítky takto odlišovat.
- Pokud transkripce bude schopna takového rozlišení, tyto informace dodavatel obdrží.
- Rozlišné rozhraní pro interní osoby a pro osoby z venku

INOVATIKA s.r.o.
- Zadavatel by měl sestavit seznam metadat, která jsou pro něj zajímavá (mluvčí, osoby o nichž se v textu hovoří,…)
- Taktéž je nutné definovat nejmenší vyhledavatelnou jednotku.

ČRo
- Nejmenší jednotkou je úsek epizody konkrétního pořadu.
- Zůstává otázkou, zda uživatel bude chtít vyhledat pouze větu nebo celý pořad, ve kterém určitá věta zazněla…
- Odkaz na analytiku vysílání mluveného slova, ta umožní rozlišovat tyto jednotlivé podčásti/jednotky.

INOVATIKA s.r.o.
- Otázka reprízovaných pořadů?
- Otázka opakovaných přepisů?

ČRo
- Reprízy se musí shlukovat do premiéry.
- Zadavatel bude specifikovat, co je premiéra a co repríza.
- V transkripci pravděpodobně nebude zadavatel požadovat opakované přepisy.

TOVEK, spol. s r.o.
- Dotaz na harmonogram připravované veřejné zakázky?
- Dotaz na vazbu na další systémy?

ČRo
- Zadavatel předpokládá přípravu zadávacích podmínek během 02/2019 a následné vyhlášení zakázky cca 03/2019
- Zadavatel musí řádně zvážit koncepci řešení.
- Všechna data budou pro dodavatele k dispozici via API zadavatele.

3) Závěr

Předběžná tržní konzultace včetně audiozáznamu byla ukončena v 17:00 hod.

Příloha č.1 – Prezenční listina účastníků

Za správnost zápisu a vyhotovení: Bc. Eva Gottová, specialista veřejných zakázek