Stručný manuál k webové aplikaci Project Finder

Aplikace Project Finder slouží k vyhledávání dokumentů (reprezentujících projekty nebo výsledky), které jsou nejpodobnější zadanému textu, přičemž míra podobnosti vychází z podobnosti textů popisujících jednotlivé dokumenty (název, klíčová slova, abstrakt/anotace/cíle projektu). **Systém pracuje s anglickým jazykem** z důvodu snadnějšího zpracování AJ oproti ČJ. Kolekci dokumentů, nad nimiž se vyhledává, lze prostřednictvím dalších voleb blíže specifikovat.

https://dafos.tacr.cz/project_finder/

Jaké jsou parametry vyhledávání – aneb co kam vyplnit...

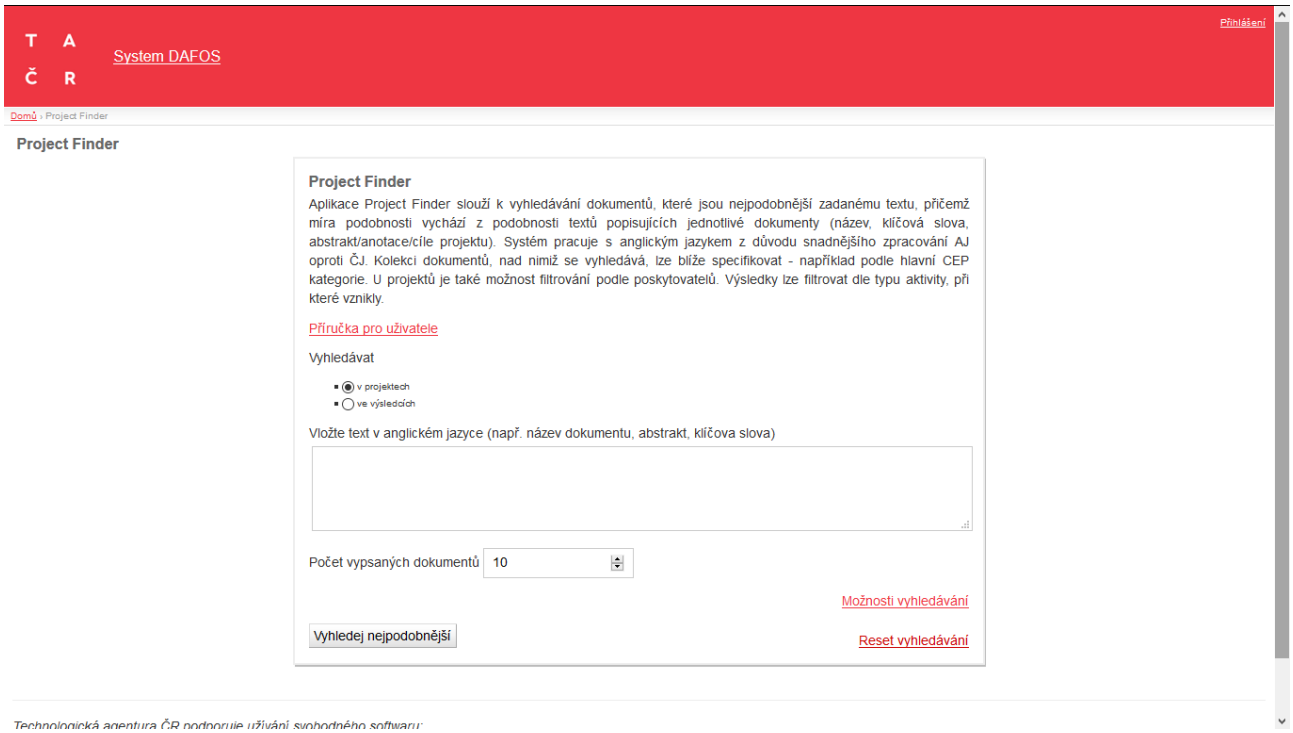
Na úvodní stránce webového rozhraní lze specifikovat následující parametry vyhledávání:

•**v jakých dokumentech se má vyhledávat:** volba (radio-button), zda vyhledávat v projektech či výsledcích (článcích, konferenčních příspěvcích, patentech, …)

•**text či seznam klíčových slov,** podle kterého/kterých se bude vyhledávat

•**počet vypisovaných dokumentů** (defaultně 10)

Screenshot úvodní stránky vyhledávání v Project Finderu



Možnosti vyhledávání – v případě projektů

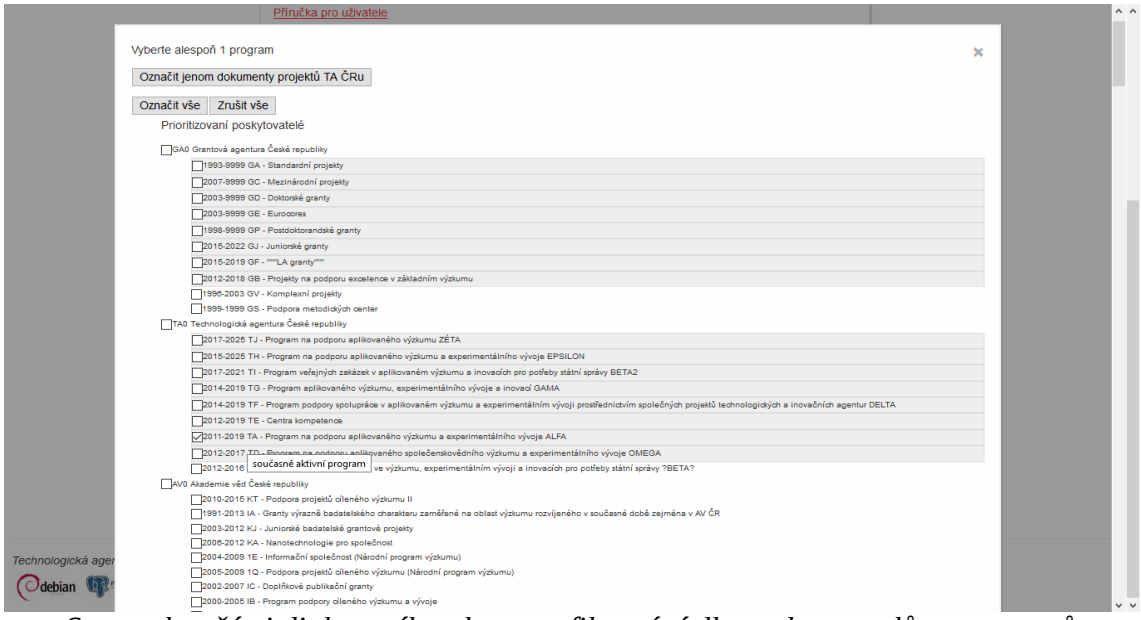
Možnosti vyhledávání se rozbalí po kliknutí na příslušný odkaz. V rozbaleném menu lze volit následující parametry:

◦ **Jaký algoritmus vyhledávání/rankování (čili uspořádání výsledků) se má použít** (algoritmus: BM25F či PG-TF) – viz poznámka níže.

◦ **Filtrování výsledků podle instituce,** která se podílela/podílí na projektu: políčko je vybaveno „našeptávačem“, který nabízí doplňování názvů institucí po zadání prvních písmen – pozor, fakulty, resp. podřízené instituce jsou k dispozici *po zadání prvních písmen nadřazené instituce*. Např. v případě Fakulty stavební ČVUT zadávejte „České vysoké učení technické v Praze / Fakulta stavební“. V případě zadávání prvních číslic IČ se zobrazují instituce, které mají IČ začínající zadanou posloupností číslic. Zvolíte-li více institucí, použije se mezi nimi logický operátor AND nebo OR zvolený v roletkovém menu (v případě, že je zvoleno AND, musí se ve vyhledaných projektech vyskytnout *všechny* zvolené instituce jako příjemci naráz, v případě OR se ve vyhledaných projektech musí vyskytnout *alespoň jedna* ze zvolených institucí).

◦ **Filtrování podle poskytovatelů a programů** – v nově otevřeném dialogovém okně je k dispozici stromová struktura pro výběr poskytovatelů a programů – zde můžete zašktnout, že Vás zajímají např. „Ministerstvo průmyslu a obchodu / TIP“, „Technologická agentura ČR / Alfa“, po volbě poskytovatelů / programů klikněte v pravém horním rohu na zavírací křížek.

◦ **Filtrování podle CEP kategorií** (oborů) – v nově otevřeném dialogovém okně je k dispozici stromová struktura pro výběr „CEP kategorií“ (číselník oborů RVVI), např. „Molekulární biologie a genetika“, „Využití počítačů“ atp. Po výběru oborů klikněte v pravém horním rohu na zavírací křížek – vyhledané projekty pak budou odpovídat alespoň jednomu z vybraných oborů.



Screenshot části dialogového okna pro filtrování dle poskytovatelů a programů

Možnosti vyhledávání – v případě projektů

Možnosti vyhledávání se rozbalí po kliknutí na příslušný odkaz. V rozbaleném menu lze volit následující parametry:

◦ **Jaký algoritmus vyhledávání se má použít** (algoritmus: BM25F či PG-TF) – viz poznámka níže)

◦ **Typ aktivity či projektu, v rámci kterého výsledek vznikl,** např.: Projekt operačního programu.

◦ **Filtrování podle CEP kategorií** (oborů) – v nově otevřeném dialogovém okně je k dispozici stromová struktura pro výběr „CEP kategorií“ (číselník oborů RVVI), např. „Molekulární biologie a genetika“, „Využití počítačů“ atp. Po výběru oborů klikněte v pravém horním rohu na zavírací křížek – vyhledané výsledky pak budou odpovídat alespoň jednomu z vybraných oborů.

Praktická ukázka (příklad s projektem na *Open Data*)

Předpokládejme, že bychom chtěli vyhledat nejpodobnější projekty k projektu s názvem „Open data in culture“, klíčovými slovy „public sector open data linked data“ a cíli, které začínají „This project is focused on preparation of open data of culture and the entire publishing process“.

Tyto texty vložíme do prvního formulářového pole (nadepsaného „Vložte text v anglickém jazyce (např. název dokumentu, abstrakt, klíčova slova) “) – interpunkce nehraje roli, velká a malá písmena rovněž ne. Zvolíme následně počet nejpodobnějších projektů, které se mají vyhledat poskytovatele (případně upřesníme programy jednotlivých poskytovatelů). Nyní již jen stačí kliknout na tlačítko „Vyhledej nejpodobnější“ a pod formulářem se následně zobrazí seznam nejpodobnějších projektů řazený sestupně dle míry podobnosti.

Po rozliknutí detailu projektu se dostanete na kartu projektu v rámci systému Dafos. Vpravo od tabulky mají uživatelé možnost zadat hodocení relevance výstupů (zda je daný výstup, čili projekt nebo výsledek je skutečně podobný zadání, resp. do jaké míry). V případě zadání jakéhokoliv hodnocení se následně v pravém dolním rohu objeví tlačítko vyzývající k zaslání hodnocení. Toto hodnocení bude dále použito při vylepšování vyhledávacích (rankovacích) algoritmů.

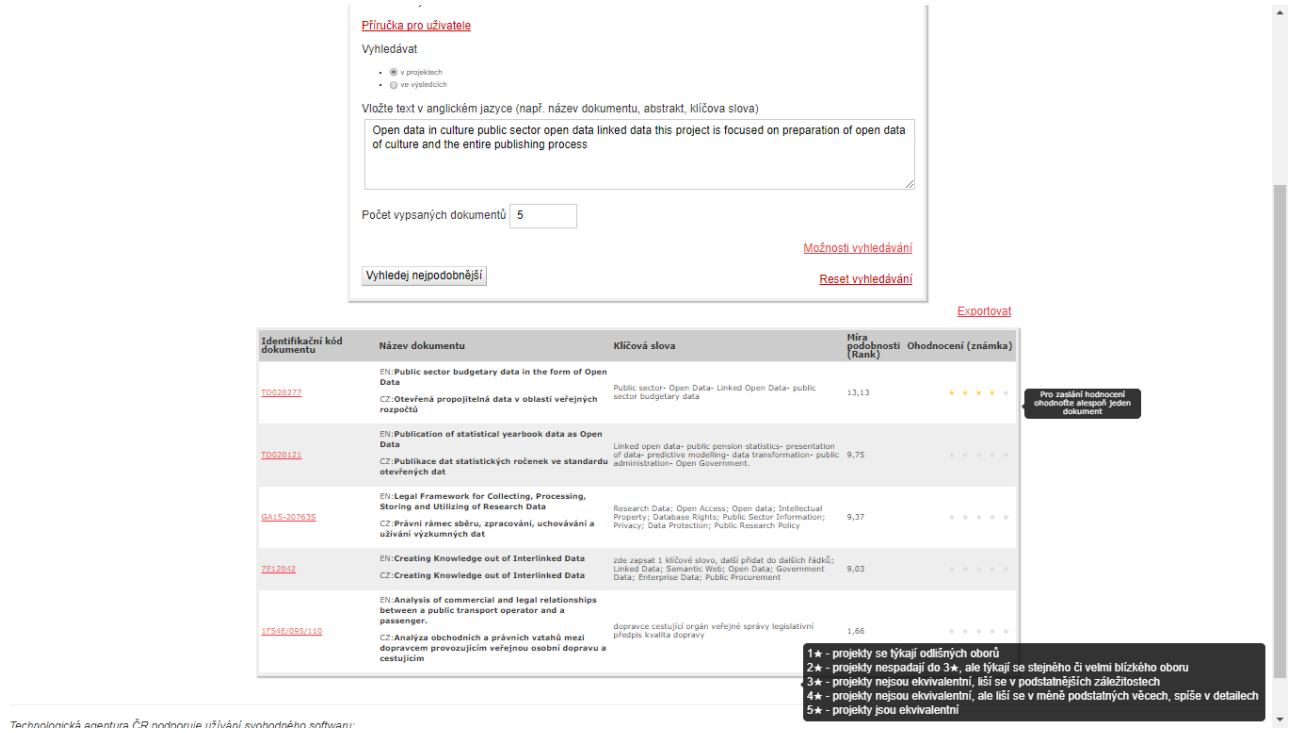
Jelikož se výsledky v závislosti na zadaných parametrech mohou lišit (např. na zvolených klíčových slovech, na zvoleném algoritmu aj.), doporučujeme experimentovat se zadanými vstupy např. takto:

> přidávat případná další klíčová slova, prodlužovat či zkracovat anotaci/cíle projektu, (spíše) vypouštět obecná slova, jako „project“, „research“, … změnit algoritmus vyhledávání/rankování.

Pozn.: PG-TF je algoritmus, který je používán v rámci SQL serveru (PostgreSQL) a který slouží k rankování (uspořádávání) textů v databázi vzhledem k dotazu na základě míry podobnosti. Více o algoritmu na stránce: https://www.postgresql.org/docs/9.1/static/textsearch-controls.html

BM25F je algoritmus pracující na podobném principu, nicméně výpočet ranku vzhledem k dotazu na základě míry podobnosti má mírně odlišné parametry. Více o algoritmu na stránce: http://www.minerazzi.com/tutorials/bm25f-model-tutorial.pdf. Z praktického hlediska však není potřeba znát principy těchto algoritmů, pouze je zkoušet.

Screenshot výsledků vyhledávání z našeho příkladu



Další aspekty týkající se vyhledávání

•Na velikosti písmen nezáleží.

•Nezáleží též na interpunkci.

•Algoritmy zatím nedokážou pracovat se synonymií (připravuje se).

•Přestože se algoritmy pokoušejí dávat nízký význam obecným slovům, které se vyskytují ve většině dokumentů, např. „the“, „and“ aj. (nebo je přímo vynechávají), je vhodnější je vynechávat.

•Vyhledávací / rankovací algoritmus se může dopouštět dvou druhů chyb: jednak že mezi vypsanými projekty se „vysoko“ objeví nerelevantní projekt, dále pak že relevantní projekt se ve výpisu neobjeví. Doporučujeme proto zkoušet různé vstupy a algoritmy. Obecně platí, že jsou preferovány na vstupu spíše kratší texty.

•Míra podobnosti vyjádřená číselně slouží pouze k porovnávání / řazení, nemá žádnou přirozenou interpretaci (ve smyslu projekt s podobností 0.5 je dvakrát podobnější než ten s podobností 0.25)

Zdroj dat, nad nimiž aplikace pracuje

Aplikace pracuje nad daty pocházejícími z Informačního systému výzkumu, experimentálního vývoje a inovací (http://www.rvvi.cz), konkrétně z části CEP a RIV. Tato data byla dále interně zpracovávána.

Sběr podnětů a zkušeností s prací s aplikací Project Finder

Budeme rádi, když nám napíšete své zkušenosti s touto aplikací a poskytete tím podněty pro další vývoj aplikace. Podněty a připomínky posílejte, prosím, na adresu: dafos@tacr.cz