



Техническа спецификация на предвидените за закупуване ДНА

I. Минимални технически и/или функционални характеристики (основание за допустимост на офертата):

Предвидената за придобиване по настоящата процедура Работна система за събиране на данни от онлайн източници по своята същност следва да представлява производствена технология, чрез която автоматизирано ще се осъществява обхождане на структурата на html източници и събиране на данни, и посредством която служителите на бенефициента ще разработват полесно скриптове и алгоритми за извличане на данни (в работна среда на Питон или еквивалентна) от html онлайн източници, екстракция на данните, техният запис, структуриране и индексирание.

Работната система за събиране на данни от онлайн източници трябва да разполага с:

1. Наличие на контролен панел* за анализ с възможност за графично представяне на общия брой свалени данни**, броя на свалените данни по източник***, броя на добавените нови източници. Филтриране**** на резултатите по дата, тип източник, проект и т.н.
2. Възможност за анализ на качеството чрез проследяване на върнатите грешки от колекторите на данни и причината в реално време.
3. Наличие на система за известяване и засичане на аномалии с възможност за:
 - a. предварително задаване на параметри на събраните данни – обем на данните, тип на данните, задължителни полета, структура и формат на данните.
 - b. Известяване по имейл.
 - c. Визуализация на аномалиите: локализация и вид на аномалията.
 - d. Манипулации през графичен интерфейс: опции за отхвърляне, редактиране, назначаване на аномалията за изследване и коригиране към екип от инженери.

Пояснение:

*) Под “контролен панел” се разбира администраторски интерфейс, който предоставя достъп до общия брой свалени данни, броя на свалените данни по източник, броя на добавените нови източници и дава възможност за филтриране на резултатите по дата, тип източник, проект и н.;

**) Под „общия брой свалени данни“ следва да се разбира броят на документите, извлечени от Работната система за събиране на данни, от всички онлайн източници;



ЕВРОПЕЙСКИ СЪЮЗ
ЕВРОПЕЙСКИ ФОНД ЗА
РЕГИОНАЛНО РАЗВИТИЕ



ОПЕРАТИВНА ПРОГРАМА
ИНОВАЦИИ И
КОНКУРЕНТОСПОСОБНОСТ

***) Под „броя на свалените данни по източник“ следва да се разбира броят на документите извлечени от конкретен източник;

****) Под „Филтриране на резултатите по дата“ следва да се разбира, че броят на показваните документи и източници трябва да има възможност да се филтрира по период;

****) Под „Филтриране на резултатите по тип източник“ следва да се разбира, че трябва да има възможност източниците да се филтрират по тип, например медия, блог, форум и пр.;

****) Под „Филтриране на резултатите по проект“ следва да се разбира, че трябва да има възможност източниците да бъдат групирани в отделни проекти;

За да се осигури посочената в т.2 по-горе възможност за анализ на качеството чрез проследяване на върнатите грешки от колекторите на данни и причината в реално време, то през администраторския интерфейс (контролния панел) следва да се осигури достъп до логове с евентуални грешки, получени в процеса на извличане на данни от онлайн източници;

За да се осигури коректното изпълнение на изискването по т.3 по-горе, а именно наличие на система за известяване и засичане на аномалии, същата следва да притежава посочените по-горе функционалности по предварително задаване на параметри на събраните данни, известяване по имейл, визуализация на аномалиите и различни манипулации, които да се извършват през графичен интерфейс. Следва да се има предвид, че посочените по-горе изисквания ще се считат изпълнени, ако:

а) Функционалността по Предварително задаване на параметри на събраните данни по отношение на параметъра „обем на данните“ дава възможност потребителят да може да задава долна и горна граница на обема от данни, при преминаването на които да получава известие по имейл;

б) Функционалността по Предварително задаване на параметри на събраните данни по отношение на параметъра „Задължителни полета“ дава възможност потребителят да може да маркира полета/елементи от събираните документи като задължителни, например задължително да присъства „заглавие“, при липсата на което потребителят трябва да получи известие по имейл;

II. Допълнителни технически и/или функционални характеристики (подлежащи на оценка)

1. Администраторският интерфейс (контролният панел) поддържа система от акаунти, през които се управляват създаването и промените по източниците и се следи кой акаунт работи по даден източник. Трябва да се поддържат минимум 3 акаунта;

www.eufunds.bg

Проект „Създаване на лабораторен център за събиране и поддържане на глобални бази данни“, договор № BG16RFOP002-1.016-0006-C01 на Сдружение за знание-интензивно развитие, финансиран от Оперативна програма „Иновации и конкурентоспособност“, съфинансирана от Европейския съюз чрез Европейския фонд за регионално развитие.



2. Предложеното от кандидата в съответствие с изискванията на т.1 от Минималните технически и/или функционални характеристики “графично представяне” е интерактивно - т.е. позволява чрез манипулация на самите графики да се извършва и филтрация по добавени източници, период за представяне и сх.
3. Възможност в администраторския интерфейс (контролния панел) потребителят да добавя и обогатява категоризацията на източниците чрез задаване на:
 - a. Име
 - b. Линк
 - c. Тип
 - d. Държава
 - e. Език
 - f. Обхват
 - g. Тема
 - h. Индустрия
 - i. Филтриране и търсене по горе изброените категории и техните стойности
 - j. Автоматично създаден уникален идентификатор.
4. Документите, извлечени от онлайн източниците, които наследяват елементите от източника (посочени в б.б. „a“, „b“, „c“, „d“, „e“, „f“, „g“, и „h“ на предходната т.3), имат възможност да съхраняват И следните елементи (при наличие):
 - a. Заглавие
 - b. Съдържание
 - c. Автор
 - d. Дата на публикуване
 - e. Линк към оригинален документ/статия, както се намира в онлайн източника.
5. Възможност за автоматично анализиране на документите чрез скрипт за разпознаване на езика, на който е написан документът. В случай, че този език е различен от езика, наследен от източника, като приоритетен се записва езикът на документа.
6. Възможност в администраторския интерфейс (контролния панел) за преименуване, добавяне на нови и премахване на таксоните и техните стойности за категоризация на източниците.
7. Възможност автоматизирания скрипт за събиране на данни да разпознава .pdf документи и да ги изпраща към трансформационен модул, който го преобразува в текстови формат.
8. Възможност на работната система за събиране на метаданни за интеракции (харесвания, споделяния и пр.) към статии, споделени в социалните мрежи.

Този документ е създаден с финансовата подкрепа на Оперативна програма „Иновации и конкурентоспособност“ 2014-2020, съфинансирана от Европейския съюз чрез Европейски фонд за регионално развитие по проект № BG16RFOP002-1.016-0006-C01, „Създаване на лабораторен център за събиране и поддържане на глобални бази данни“.

www.eufunds.bg

Проект „Създаване на лабораторен център за събиране и поддържане на глобални бази данни“, договор № BG16RFOP002-1.016-0006-C01 на Сдружение за знание-интензивно развитие, финансиран от Оперативна програма „Иновации и конкурентоспособност“, съфинансирана от Европейския съюз чрез Европейския фонд за регионално развитие.



ЕВРОПЕЙСКИ СЪЮЗ
ЕВРОПЕЙСКИ ФОНД ЗА
РЕГИОНАЛНО РАЗВИТИЕ



The Knowledge-Intensive
Development Association



ОПЕРАТИВНА ПРОГРАМА
ИНОВАЦИИ И
КОНКУРЕНТОСПОСОБНОСТ

Цялата отговорност за съдържанието на документа се носи от Сдружение за знание-интензивно развитие и при никакви обстоятелства не може да се приема, че този документ отразява официалното становище на Европейския Съюз и Управляващия орган.

www.eufunds.bg

Проект „Създаване на лабораторен център за събиране и поддържане на глобални бази данни“, договор № BG16RFOP002-1.016-0006-C01 на Сдружение за знание-интензивно развитие, финансиран от Оперативна програма „Иновации и конкурентоспособност“, съфинансирана от Европейския съюз чрез Европейския фонд за регионално развитие.