



**Avviso di ricognizione per la selezione di un'azienda fornitrice di dati elaborati
tramite sistemi di intelligenza artificiale per uno studio metodologico nell'ambito
di attività di ricerca**

**Il Direttore del Dipartimento di Statistica, Informatica, Applicazioni
"Giuseppe Parenti" (DISIA) dell'Università degli Studi di Firenze**

- **CONSIDERATO** che il Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti" (DiSIA) ha interesse ad attivare una selezione per individuare un'azienda disponibile a fornire a titolo gratuito, dati utili alla ricerca attraverso l'impiego di soluzioni basate sull'intelligenza artificiale;
- **VISTO** il DPIA (Data Protection Impact Assessment), redatto ai sensi dell'articolo 35 del GDPR, con cui è stato valutato l'impatto sulla protezione dei dati;
- **VISTA** la delibera n. 221 del 18/12/2024 con cui il Consiglio di Dipartimento di Statistica, Informatica, Applicazioni "Giuseppe Parenti" (DiSIA) ha approvato la suddetta richiesta di avviso di ricognizione;
- **CONSIDERATA** l'urgenza di acquisire i dati necessari per consentire l'avvio della ricerca nel più breve tempo possibile;
- **RILEVATO** che la tempestiva disponibilità dei dati costituisce un presupposto imprescindibile per l'avvio e il regolare svolgimento dell'attività di ricerca e che un eventuale ritardo nella loro acquisizione potrebbe compromettere il rispetto delle tempistiche progettuali e degli obiettivi scientifici prefissati;

EMANA IL SEGUENTE AVVISO DI RICOGNIZIONE

Articolo 1 – Oggetto dell'avviso

il presente avviso è finalizzato a individuare un'azienda in grado di fornire a titolo gratuito, al Dipartimento di Statistica Informatica, Applicazioni "Giuseppe Parenti" (DiSIA) dell'Università degli Studi di Firenze, una serie di dati elaborati mediante sistemi di intelligenza artificiale, necessari per lo svolgimento di uno studio metodologico. Nell'ambito della ricerca, potranno essere realizzate pubblicazioni scientifiche e organizzati eventi divulgativi dei risultati raggiunti, a cui l'azienda potrà essere invitata a collaborare e a partecipare.

Articolo 2 – Requisiti di partecipazione

Possono manifestare interesse le aziende che soddisfano i seguenti requisiti:

- a. Esperienza comprovata nell'ambito del webscraping e webcrawling su larga scala da siti web aziendali e social media (e.g., LinkedIn).
- b. Capacità di utilizzo di tecniche avanzate di elaborazione del linguaggio naturale (NLP), inclusi modelli linguistici di grandi dimensioni (LLM).
- c. Esperienza comprovata nell'analisi testuale tramite conteggio di keyword e metodi di similarità semantica.
- d. Esperienza comprovata nell'operare su dataset specifici, principalmente il registro ORBIS.

Art. 3 Attività da svolgere per il progetto

Il progetto di ricerca *"Integrazione di dati provenienti da webscraping con dati tradizionali per produrre stime di popolazione robuste sulla sensibilità delle aziende italiane all'innovazione e alla sostenibilità"* si propone di investigare e approfondire metodi statistici per l'integrazione di dati provenienti da fonti non probabilistiche e fonti probabilistiche. In particolare, il progetto è volto all'utilizzo di tali metodi per produrre stime relative alla sensibilità delle aziende italiane in merito ai temi legati all'innovazione e alla sostenibilità. L'interesse verso tali tematiche è legato alla scarsa disponibilità di dati provenienti da fonti tradizionali di tipo probabilistico. Lo sviluppo del progetto richiede pertanto la disponibilità di dati innovativi, intrinsecamente non probabilistici, che permettano di disporre in tempi rapidi delle informazioni necessarie. I dati provenienti da webscraping e webcrawling potrebbero soddisfare tale esigenza.

I dati non probabilistici, ottenuti tramite selezione arbitraria con probabilità di inclusione sconosciute, rendono inapplicabili i metodi di inferenza del campionamento probabilistico e introducono bias di selezione. Di qui la necessità per fare inferenza mediante i dati non probabilistici e di integrarli con dati esterni quali potrebbero essere quelli provenienti da campione probabilistici. Per poter integrare i dati innovativi con quelli tradizionali è necessario non solo avere variabili ausiliare comuni alle due fonti di dati, ma anche che le due fonti siano riferite alla stessa popolazione di studio. Pertanto, i dati provenienti da webscraping devono fare riferimento ad una popolazione per cui esistono delle fonti probabilistiche e

quindi ad una lista di aziende quale potrebbe essere il registro ORBIS¹. Inoltre, la ricerca potrebbe essere arricchita se i dati testuali provenienti dai siti web delle aziende fossero integrati da eventuali dati raccolti sui profili social (e.g., LinkedIn) appartenenti alle stesse. I dati raccolti attraverso le operazioni di webscraping non sono utilizzabili direttamente nello studio, ma richiedono di essere rielaborati mediante l'utilizzo di modelli linguistici di grandi dimensioni (LLM) e metodi di classificazione al fine di creare delle variabili e indicatori specifici che misurino la sensibilità delle aziende all'innovazione e sostenibilità.

Pertanto le attività richieste all'azienda sono le seguenti:

- Raccolta e rielaborazione dei dati di webscraping su aziende appartenenti ad una specifica popolazione quali le aziende toscane presenti registro ORBIS.
- Rielaborazione dei dati testuali raccolti al fine di costruire variabili e indicatori sulla sensibilità delle aziende all'innovazione e alla sostenibilità.
- Collaborazione con i ricercatori del DiSIA coinvolti nel progetto per la definizione/integrazione delle variabili e indicatori sulla sensibilità delle aziende all'innovazione e alla sostenibilità di cui al punto precedente.
- Fornitura dei dati già rielaborati (trasformati nelle variabili come sopra specificato) entro 60 giorni dall'inizio della collaborazione.
- Disponibilità ad offrire un eventuale supporto/assistenza relativamente all'utilizzo dei dati nella fase successiva di analisi/utilizzo degli stessi da parte dei ricercatori del DiSIA per le finalità del progetto specificate sopra.

Art. 3.1 – Caratteristiche dei dati e tempi di consegna

L'obiettivo principale del progetto è la produzione di stime a livello aggregato, escludendo qualsiasi analisi su dati individuali. Le metodologie adottate garantiranno che le informazioni trattate siano finalizzate esclusivamente alla generazione di statistiche di sintesi, nel pieno rispetto della riservatezza e della protezione dei dati personali.

Pertanto, si richiede che i dati raccolti dall'azienda fornitrice tramite tecniche di webscraping non siano forniti nella loro forma grezza, ma solo dopo un processo di elaborazione che ne assicuri l'aggregazione. In particolare, dovranno essere trasmesse esclusivamente variabili derivate da analisi testuali basate su parole

¹ banca dati di Bureau Van Dijk. Le banche dati Bureau van Dijk, società acquisita da Moody's Analytics nel 2017, raccolgono i dati finanziari delle imprese e consentono la selezione di campioni di imprese per settore, area geografica, e dimensioni quantitative delle voci finanziarie.

chiave, evitando così che i ricercatori del DiSIA possano accedere, memorizzare o utilizzare direttamente contenuti testuali originali. Questo approccio consentirà loro di estrarre informazioni di sintesi senza trattare dati personali o informazioni identificative.

I dati raccolti tramite webscraping rappresentano un campione non probabilistico di aziende. Per poterli utilizzare a fini inferenziali sulla popolazione delle aziende toscane, è necessario integrarli con dati provenienti da fonti censuarie o da campioni probabilistici selezionati dalla stessa popolazione. Di conseguenza, l'azienda fornitrice dovrà garantire che i dati trasmessi riguardino un sottoinsieme di una specifica popolazione di aziende, come quelle incluse negli archivi ORBIS o AIDA. Inoltre, per ciascuna azienda, dovrà fornire oltre alle variabili rielaborate derivate dal webscraping anche variabili ausiliarie di natura esclusivamente socioeconomica, demografica e geografica, necessarie al progetto per migliorare la qualità delle stime aggregate. Queste informazioni dovranno essere estratte da registri ufficiali (ad esempio ORBIS e AIDA) prima di consegnare il dataset, rimuovendo qualsiasi dato identificativo sensibile (come partita IVA, ragione sociale o indirizzo). L'uso di tali variabili sarà finalizzato esclusivamente al miglioramento della qualità delle stime aggregate, senza alcuna possibilità di identificare singole aziende o individui.

L'azienda aggiudicataria dovrà consegnare i dati richiesti, come specificato nei dettagli del bando, entro 60 giorni dalla data di affidamento dell'incarico. Le modalità operative, le tempistiche intermedie e gli eventuali step di consegna potranno essere concordati con la Responsabile del progetto, Prof.ssa Emilia Rocco.

Articolo 4 – Modalità e termini di presentazione della manifestazione di interesse

Le aziende interessate dovranno compilare la domanda di partecipazione (conforme al modello allegato 1, formato PFD/A) sottoscritta con firma digitale (o in alternativa con scansione del modello compilato e firmato unitamente a copia del documento di identità del legale rappresentante firmatario). La manifestazione dovrà includere:

- a) una presentazione dell'azienda e delle sue attività;
- b) una descrizione dettagliata dell'esperienza, in termini di progetti realizzati, nella raccolta di dati tramite webscraping e webcrawling su larga scala da siti web aziendali e social media e nell'elaborazione dei dati di tipo testuale;
- c) una proposta progettuale che descriva dettagliatamente i dati che l'azienda può rendere disponibili e che includa un piano dettagliato delle attività con una chiara

definizione delle tempistiche e delle fasi del processo di raccolta ed elaborazione dati, specificando:

- la tipologia di dati acquisibili tramite webscraping e webcrawling (e.g., contenuti testuali da siti web aziendali, dati estratti da social media come LinkedIn, altre fonti rilevanti);
- la loro pertinenza rispetto agli obiettivi dello studio, con particolare riferimento alla misurazione della sensibilità delle aziende italiane ai temi dell'innovazione e della sostenibilità;
- le metodologie e gli strumenti utilizzati per la raccolta e il trattamento dei dati, inclusi eventuali modelli NLP e LLM, tecniche di classificazione, analisi semantica.

La domanda e la relativa documentazione allegata dovrà essere inviata mediante posta elettronica certificata (PEC) all'indirizzo disia@pec.unifi.it, indicando come oggetto: DiSIA - "Bando per raccolta dati- candidatura nome azienda", entro e non oltre le ore 13:30 del giorno 01/04/2025, pena l'esclusione dalla procedura valutativa.

I dati trasmessi con la domanda di partecipazione al progetto saranno trattati per le finalità di gestione della procedura di selezione. A tal fine, le imprese candidate sono invitate a prendere visione dell'informativa per il trattamento dei dati ai sensi del GDPR 2016/679, consultabile all'indirizzo <https://www.unifi.it/p11360.html>.

Il Dipartimento di Statistica, Informatica, Applicazioni non assume alcuna responsabilità per eventuali disguidi telematici o comunque imputabili a fatto di terzi, a caso fortuito o forza maggiore.

Le candidature sono ammesse alla valutazione comparativa con riserva e in ogni momento ne può essere disposta l'esclusione, con provvedimento motivato per difetto di requisiti da parte del Direttore del Dipartimento.

Articolo 5 - Obblighi di riservatezza e tutela della ricerca

L'azienda sarà tenuta a rispettare gli obblighi di non concorrenza e riservatezza e si asterrà da ogni attività che possa generare conflitto di interesse con le finalità del progetto, adotterà altresì tutte le misure necessarie per garantire che la riservatezza e la sicurezza delle informazioni saranno rispettate dai propri dipendenti o da qualsiasi altra persona che agisca per suo conto e che possa avere l'opportunità di acquisire informazioni relative al Progetto di ricerca. I risultati della ricerca saranno di proprietà dell'Università degli Studi di Firenze, che avrà il diritto di pubblicare liberamente i risultati delle attività svolte nell'ambito del Progetto di Ricerca, e potranno essere condivisi con tutti i partner del progetto. Le informazioni personali relative all'azienda partecipante saranno raccolte, tenute e gestite unicamente dal

dipartimento DiSIA dell'Università degli Studi di Firenze, nel rispetto dei principi di liceità, correttezza, trasparenza, adeguatezza, pertinenza e necessità di cui all'articolo 5, comma 1 del Regolamento Generale sulla Protezione dei Dati (GDPR). I dati saranno resi accessibili solo a coloro che, sia all'interno progetto, che all'esterno, ne abbiano bisogno esclusivamente per la gestione delle attività necessarie al suo svolgimento.

Il responsabile scientifico delle attività di ricerca è la prof.ssa Emilia Rocco - mail emilia.rocco@unifi.it

Articolo 6 – Valutazione delle manifestazioni di interesse

Le manifestazioni di interesse pervenute saranno valutate da una commissione appositamente nominata dal Direttore del Dipartimento di Statistica, Informatica, Applicazioni, sulla base dei seguenti criteri:

- a) esperienza e competenza dell'azienda nel settore con particolare riferimento ai temi del progetto;
- b) qualità, pertinenza e tempestività della proposta progettuale.

La Commissione giudicatrice avrà a disposizione 50 punti totali, così distribuiti:

- a) esperienza comprovata nelle specifiche attività di interesse (webscraping, elaborazioni dati testuali, familiarità con il registro ORBIS) 10 punti;
- b) valutazione della proposta progettuale presentata (qualità, tempestività e adeguatezza degli strumenti proposti) 40 punti.

Articolo 7 – Trattamento dei dati personali

I dati raccolti sono trattati e conservati ai sensi del Regolamento UE n. 2016/679 relativo alla protezione delle persone fisiche con riguardo al trattamento dei dati personali, nonché alla libera circolazione di tali dati, del decreto legislativo 30 giugno 2003, n.196 recante il "Codice in materia di protezione dei dati personali" e ss mm e ii.

I dati personali raccolti nell'ambito della presente procedura, saranno trattati esclusivamente per lo svolgimento delle attività amministrative e per l'assolvimento degli obblighi previsti dalle leggi e dai regolamenti aziendali in materia.

Il titolare del trattamento dei dati personali è l'Università degli Studi di Firenze, con sede in Firenze, Piazza San Marco, 4 telefono 055 27571 e-mail: urp@unifi.it, PEC ateneo@pec.unifi.it. I dati di contatto del Responsabile per la protezione dei dati sono email: privacy@adm.unifi.it PEC: protezionedati@pec.unifi.it. I rapporti per la protezione dei dati derivanti dall'oggetto del presente avviso di cui all'art. 1 tra

l'Università degli Studi di Firenze e l'aggiudicatario saranno disciplinati con atto specifico.

I partecipanti all'avviso sono pregati di prendere visione dell'informativa per il trattamento dei dati personali di operatori economici (o loro legali rappresentanti) interessati a partecipare a procedure di scelta del contraente, fornitori di beni e servizi, fornitori di attività di job placement e tirocinio, collaboratori esterni pubblicata al seguente link:

https://www.unifi.it/sites/default/files/migrated/documents/Informativa_TERZI.pdf

Articolo 8 - Pubblicità e avvio delle attività

La graduatoria del presente avviso sarà pubblicata nell'Albo ufficiale dell'Ateneo all'indirizzo <https://www.unifi.it/it/ateneo/albo-ufficiale> e sul sito web del Dipartimento <https://www.disia.unifi.it/vp-117-bandi-e-avvisi.html>. Successivamente l'azienda vincitrice verrà contattata per concordare l'inizio delle attività.

Articolo 9 – Disposizioni Finali

Il presente avviso è finalizzato alla selezione di un operatore economico che si impegnerà a fornire i dati richiesti secondo le modalità descritte.

La fornitura dei dati avverrà a titolo completamente gratuito e non comporterà alcun obbligo economico per il Dipartimento. La società selezionata non potrà avanzare richieste di compenso, rimborso spese o altra forma di corrispettivo economico per la prestazione fornita.

La selezione della società non determina la costituzione di un rapporto contrattuale né impegna il Dipartimento alla stipula di successivi accordi di collaborazione o affidamenti di servizi.

Il Dipartimento si riserva la facoltà di modificare, sospendere o interrompere la presente procedura in qualsiasi momento, senza che ciò comporti alcun obbligo risarcitorio nei confronti degli operatori economici partecipanti.

Articolo 10 - Foro competente

Qualunque controversia in ordine all'interpretazione ed esecuzione del successivo affidamento in merito al presente avviso sarà devoluta alla esclusiva competenza dell'Autorità Giudiziaria del Foro di Firenze



Articolo 11 - Unità organizzativa competente e responsabile del procedimento

Ai sensi della Legge n. 241/1990 e s.m. nonché del "Regolamento sui procedimenti amministrativi dell'Università degli Studi di Firenze" è individuata quale unità organizzativa competente il Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Viale G.B. Morgagni, n. 59 - Firenze, tel. 055/2751578-1580-1582, e-mail: disia@disia.unifi.it, PEC disia@pec.unifi.it.

Il responsabile del procedimento è la Dott.ssa Claudia Conti.

**Il Direttore del Dipartimento
Prof.ssa Emanuela Dreassi**