PROCEDURA APERTA SOPRA SOGLIA COMUNITARIA PER L'AFFIDAMENTO DEL SERVIZIO "DESIGN AND IMPLEMENTATION OF METHODOLOGIES FOR THE CONSTRUCTION OF KNOWLEDGE GRAPHS WITH (I) FAIR PRINCIPLES AND (II) LINKED DATA PARADIGM" CON IL CRITERIO DELL'OFFERTA ECONOMICAMENTE PIÙ VANTAGGIOSA SULLA BASE DEL MIGLIOR RAPPORTO QUALITÀ/PREZZO NELL'AMBITO DEL PIANO NAZIONALE RIPRESA E RESILIENZA (PNRR) MISSIONE 4 , "ISTRUZIONE E RICERCA" COMPONENTE 2, "DALLA RICERCA ALL'IMPRESA" INVESTIMENTO 3.1, "FONDO PER LA REALIZZAZIONE DI UN SISTEMA INTEGRATO DI INFRASTRUTTURE DI RICERCA E INNOVAZIONE" PROGETTO FOSSR CUP B83C22003950001 CUI S80054330586202300112 CIG A038EC14C3

CAPITOLATO SPECIALE DI APPALTO

- Parte Tecnica -

1. PREMESSE 3
2. CARATTERISTICHE TECNICHE/FUNZIONALITÀ E DOTAZIONI MINIME DEL SERVIZIO/DELLA FORNITURA 4
2.1. CARATTERISTICHE GENERALI DELLA FORNITURA 4
2.2. CARATTERISTICHE DEL SISTEMA 5
2.2.1. METODO (E SUA IMPLEMENTAZIONE) PER LA GENERAZIONE E CLASSIFICAZIONE DI COMPETENCY QUESTION A PARTIRE DA USER STORY E DA DATASET ESISTENTI 5
2.2.2. SISTEMA DI GENERAZIONE DI ONTOLOGIE A PARTIRE DA COMPETENCY QUESTION 5
2.2.3. SISTEMA DI VALIDAZIONE DI KNOWLEDGE GRAPH 6
2.2.4. SISTEMA DI GENERAZIONE DI ONTOLOGIE E KNOWLEDGE GRAPH A PARTIRE DA DATASET 6
2.2.5. FORMAZIONE 6
2.2.6. SCADENZE E FASI DI IMPLEMENTAZIONE 7
2.2.7. GARANZIA 7
2.2.8. ASSISTENZA TECNICA, SUPPORTO E MANUTENZIONE 8
3. MODALITÀ DI REALIZZAZIONE DEL SERVIZIO 8
3.1. PIANO DI PROGETTO 9

1. PREMESSE

La Stazione Appaltante, Istituto di Scienze e Tecnologie della Cognizione del Consiglio Nazionale delle Ricerche (CNR-ISTC), sede centrale di Roma, intende procedere mediante procedura di gara all'affidamento di un servizio di progettazione, sviluppo, installazione, test e manutenzione di una piattaforma di software modulare che implementi ed estenda la metodologia di modellazione di grafi della conoscenza eXtreme Design (cf Figura 1) con strumenti di deep learning e large language model finalizzati alla costruzione di grafi della conoscenza semantici a partire da sorgenti eterogenee strutturate (ad es. database relazionali), semi-strutturate (ad as. CSV) e non strutturate (ad es. linguaggio naturale). I grafi della conoscenza risultanti dovranno essere conformi con

1. la rete di ontologie FOSSR;
2. i principi FAIR;
3. il paradigma dei linked open data.

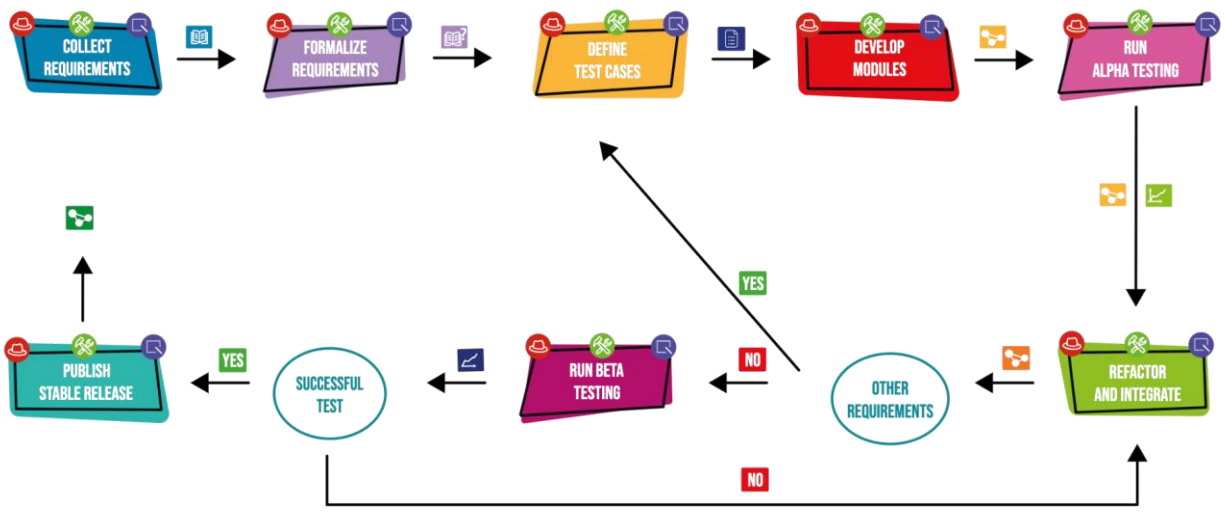

Figura 1. La metodologia eXtreme Design.

La piattaforma risultante permetterà

- l'integrazione con sistemi di estrazione della conoscenza finalizzati
  - alla generazione di ontologie a partire da requisiti espressi mediante competency question o scenari d'uso del knowledge graph in linguaggio naturale;
  - al popolamento automatico di un grafo della conoscenza a partire da corpora testuali.
- la validazione dei grafi della conoscenza attraverso unit testing volto a validare aspetti strutturali, logici e funzionali dei grafi stessi
- l'integrazione con i linguaggi dichiarativi standard per la mappatura di sorgenti eterogenee strutturate (ad es. database relazionali) e semi-strutturate (ad as. CSV) fornendo
  - un motore per l'esecuzione di regole di mappatura;
  - un'interfaccia basata su applicativo Web per l'editing di regole di mappatura di facile utilizzo;
  - esperimenti sulla generazione automatica di regole di mappatura dati in input una sorgente di dati ed una rete di ontologie.

La piattaforma sarà il cuore del sistema di ingegneria della conoscenza nell'ambito del progetto "FOSSR: Fostering Open Science in Social Science Research", il cui obiettivo è la creazione di una infrastruttura di ricerca distribuita atta a offrire strumenti e servizi di supporto alla comunità scientifica nell'ambito delle scienze sociali. L'infrastruttura di ricerca che sarà realizzata coinvolgerà altre infrastrutture di ricerca esistenti coordinate dal CNR, quali CESSDA, RISIS e SHARE, oltre a usufruire di dati statistici provenienti da ISTAT. Per raggiungere tale scopo verrà progettato un Cloud Italiano per l'Open Science, seguendo le linee guida del progetto "European Open Science Cloud", in cui integrare servizi innovativi relativi alla raccolta, all'analisi e alla cura dei dati, seguendo i principi FAIR (Findable, Accessible, Interoperable, Reusable).
La piattaforma dovrà essere sviluppata e configurata avvalendosi delle apparecchiature hardware in fase di acquisizione, che saranno installate in quattro data center (nodi) dislocati in sedi differenti del CNR.

2. CARATTERISTICHE TECNICHE/FUNZIONALITÀ E DOTAZIONI MINIME DEL SERVIZIO/DELLA FORNITURA

L'offerta dell'operatore economico dovrà rispettare tutte le caratteristiche tecniche, funzionalità e dotazioni minime della fornitura stabilite nel presente paragrafo, pena l'esclusione dalla procedura di gara.
Ai sensi di quanto previsto nell'allegato II.5 del D.Lgs. 36/2023 (di seguito anche "Codice") l'offerente dimostra, nella propria offerta, con qualsiasi mezzo appropriato, compresi i mezzi di prova di cui all'articolo 105 del Codice, che le soluzioni proposte ottemperano in maniera equivalente alle prestazioni, ai requisiti funzionali e alle specifiche tecniche prescritti nel presente documento.
I seguenti sotto paragrafi andranno ad esplicitare le necessità della fornitura in termini software infrastrutturale e software applicativo.

2.1. Caratteristiche generali della fornitura

Il sistema dovrà essere studiato, progettato ed implementato nel linguaggio Python (versione >= 3.10) come un software composto da più moduli indipendenti ed interoperabili. Il grado di accoppiamento ed isolamento dei moduli sarà uno degli elementi di valutazione della piattaforma in fase di collaudo.
Ogni modulo dovrà fornire:

- interfacce di comunicazione accessibili tramite API REST su HTTP;
- software development kit (SDK) per favorire l'estensione d il riuso in altri software in ambiente Python;
- un'interfaccia grafica web-based che favorisca l'uso delle funzionalità del modulo da parte di un utente attraverso un comune web browser;
- documentazione della progettazione ed implementazione del codice;
- guida utente con esempi per l'uso del modulo attraverso le varie modalità di interazione (API REST, SDK ed interfaccia utente).

L'obiettivo del sistema sarà studiare, progettare ed implementare alcune attività della metodologia eXtreme Design[1] sfruttando tecniche di deep learning e large language model. Le specifiche tecniche di deep learning e gli opportuni large language model saranno identificati dall'operatore economico attraverso un'analisi dello stato dell'arte. Altresì il sistema dovrà essere progettato ed implementato seguendo lo stile architetturale component-based.

Il software dovrà essere rilanciato in formato open source con licenza Apache 2.0[2].

[1] Presutti V, Daga E, Gangemi A, Blomqvist E. eXtreme Design with Content Ontology Design Patterns. In: Blomqvist E, Sandkuhl K, Scharffe F, Sva´tek V, editors. Proc. of WOP 2009. vol. 516 of CEUR Workshop Proceedings. CEUR-WS.org; 2009.
[2] https://www.apache.org/licenses/LICENSE-2.0

2.2. Caratteristiche del sistema

2.2.1. Metodo (e sua implementazione) per la generazione e classificazione di competency question a partire da user story e da dataset esistenti

Il sistema dovrà fornire un modulo per produrre una lista di competency question a partire da:

- scenari d'uso del knowledge graph espressi in inglese o italiano. Uno scenario d'uso è una descrizione delle sequenze di azioni che definisce una particolare interazione tra attore e il knowledge graph che tipicamente porta alla definizione di requisiti ontologici, ossia competency question;
- dataset sia in formato strutturato che semistrutturato. Per dataset in formato strutturato si intendono dati contenuti in database relazionali, XML o simili. Al contempo, per dataset semistrutturati si intendono dati contenuti in formati come CSV, XSLX e JSON. Il numero e la tipologia di sorgenti supportate è un criterio su cui verrà valutata l'offerta tecnica.

Il sistema dovrà prendere come input uno o più dataset/scenari e produrre come output una lista di competency question che potrà essere utilizzata per successive fasi ontology engineering anche avvalendosi del sistema stesso (cf. Sezione 2.2.2). Il cuore di questo modulo dovrà essere un motore di analisi che si dovrà avvalere delle più moderne tecniche di data analysis basate su deep learning anche avvalendosi di large language model.

2.2.2. Sistema di generazione di ontologie a partire da competency question

La metodologia agile di ingegneria eXtreme Design prevede, come molte altre metodologie in letteratura, l'uso delle competency question[3] (CQ) come soluzione per rappresentare i requisiti di modellazione di grafi della conoscenza (KG). Le CQ sono domande espresse in linguaggio naturale a cui un KG deve rispondere una volta modellato.
Il sistema dovrà essere progettato ed implementato per prendere in input una lista di CQ e produrre come output un'ontologia formalizzata attraverso il linguaggio OWL versione 2. L'ontologia risultante dovrà essere modellata in modo da rispondere in maniera funzionale alle CQ attraverso il riuso di Ontology Design Patterns[4], frame linguistici di Framester[5], ontologie di OntoPiA[6] ed integrata alla rete di ontologie del progetto FOSSR esistente. Al centro del sistema ci dovrà essere un motore di elaborazione del linguaggio naturale, di estrazione della conoscenza e di produzione di artefatti ontologici basato su tecniche di deep learning anche avvalendosi di large language model. Tale sistema dovrà essere sviluppato a valle di un lavoro di analisi a carico dell'operatore economico e volto ad identificare i sistemi e le soluzioni più efficaci allo stato dell'arte per lo specifico compito, come, ad esempio, modelli generativi, large language model e graph neural network.

[3] Grüninger M, Fox MS. The role of competency questions in enterprise engineering. In: Benchmarking—Theory and practice. Springer; 1995. p. 22-31.
[4] http://ontologydesignpatterns.org/
[5] https://framester.github.io/
[6] https://github.com/italia/daf-ontologie-vocabolari-controllati

2.2.3. Sistema di validazione di knowledge graph

Il sistema dovrà essere progettato ed implementato per generare automaticamente test di unità espressi sotto forma di query SPARQL. I test di unità saranno generati dal sistema applicando tecniche di deep learning anche avvalendosi di large language model. Il sistema dovrà essere sviluppato a valle di un lavoro di analisi a carico dell'operatore economico e volto ad identificare i sistemi e le soluzioni più efficaci allo stato dell'arte per lo specifico compito

2.2.4. Sistema di generazione di ontologie e knowledge graph a partire da dataset

Il sistema dovrà fornire un modulo per generare un'ontologia prendendo come input uno o più dataset. Anche in questo caso il sistema dovrà gestire sia dataset strutturati che semistrutturati. Il numero e la tipologia di dataset supportati dal sistema sarà un elemento di valutazione dell'offerta tecnica. L'ontologia risultante dovrà formalizzata con il linguaggio OWL versione 2 e modellata attraverso il riuso di Ontology Design Patterns[7], frame linguistici di Framester[8], ontologie di OntoPiA[9] ed integrata alla rete di ontologie del progetto FOSSR esistente e a cui il sistema stesso contribuirà alla creazione.
Il sistema dovrà fornire un modulo per generare linked data prendendo come input uno o più dataset ed una rete di ontologie. Come per il caso precedente, il sistema dovrà gestire sia dataset strutturati che semistrutturati. Il numero e la tipologia di dataset supportati dal sistema sarà un elemento di valutazione dell'offerta tecnica. I linked data risultanti dovranno essere rappresentati con il metamodello RDF secondo i secondo i principi Linked Data[10] e modellati conformemente alla rete di ontologie fornita come input. Il risultato della computazione che il modulo dovrà fornire sarà un dataset in Linked Data corredato da un descrittore delle regole generate ed utilizzate dal sistema per mappare i dataset di input in linked data secondo la rete di ontologie fornita. Tali regole di mappatura dovranno essere rappresentate nel linguaggio RML[11]. Il sistema dovrà essere sviluppato a valle di un lavoro di analisi a carico dell'operatore economico e volto ad identificare i sistemi e le soluzioni più efficaci allo stato dell'arte per lo specifico compito come, ad esempio, modelli generativi e large language model

2.2.5. Formazione

L'operatore economico dovrà garantire un programma di addestramento all'uso dei software installati (sia infrastrutturali che applicativi) di durata minima effettiva di almeno 72 ore, fatta salva l'offerta migliorativa presentata in sede di gara[12]: il programma dovrà essere tenuto preferibilmente on-site presso la sede di consegna ed installazione, da personale specializzato, secondo un calendario che dovrà essere concordato con la stazione appaltante. Detto programma dovrà essere avviato entro 30 giorni solari dal superamento della verifica di conformità della strumentazione, salvo diverso accordo. Il corso e la documentazione di addestramento dovranno essere in lingua italiana e/o inglese.

[7] http://ontologydesignpatterns.org/
[8] https://framester.github.io/
[9] https://github.com/italia/daf-ontologie-vocabolari-controllati
[10] https://www.w3.org/wiki/LinkedData
[11] https://rml.io/
[12] Se nel disciplinare è presente una premialità correlata al miglioramento

2.2.6. Scadenze e fasi di implementazione

L'operatore economico dovrà fornire una dettagliata specifica delle scadenze e delle fasi di implementazione dei servizi nell'ambito del progetto. Questo requisito è finalizzato a garantire una consegna tempestiva dei servizi e una pianificazione delle attività adeguata e ben strutturata, al fine di raggiungere gli obiettivi prefissati in modo efficiente e conforme alle aspettative dell'appaltante.
Le principali caratteristiche da rispettare sono le seguenti:

- **Pianificazione dettagliata**
  L'operatore economico dovrà fornire una pianificazione dettagliata delle attività di implementazione dei servizi. Saranno specificate le fasi di sviluppo, configurazione, test e messa in produzione dei servizi, con l'indicazione delle risorse assegnate e delle tempistiche previste per ciascuna fase.
- **Scadenze e milestone**
  Saranno definite chiaramente le scadenze e le milestone (punti di controllo) chiave del progetto. Queste saranno specificate in termini di date di completamento delle attività principali e di consegna dei servizi, al fine di monitorare l'avanzamento del progetto e garantire il rispetto dei tempi concordati.
- **Documentazione delle fasi di implementazione**
  Sarà richiesta la documentazione dettagliata delle fasi di implementazione dei servizi, evidenziando le attività previste e le tempistiche per ciascuna fase. Questa documentazione consentirà una chiara comprensione delle operazioni da parte della Stazione Appaltante e delle parti coinvolte.
- **Verifica e controllo**
  L'operatore economico dovrà attuare meccanismi di verifica e controllo dell'avanzamento del progetto rispetto alla pianificazione stabilita. Saranno previsti strumenti per rilevare eventuali scostamenti rispetto ai tempi previsti e per prendere misure correttive tempestive.
- **Comunicazione con la Stazione Appaltante**
  Sarà garantita una costante comunicazione con la Stazione Appaltante riguardo all'avanzamento delle fasi di implementazione e delle scadenze. Eventuali ritardi o variazioni rispetto alla pianificazione dovranno essere tempestivamente segnalati alla Stazione Appaltante, accompagnati da una chiara spiegazione e da eventuali azioni correttive.
- **Revisione e approvazione**
  La pianificazione e le scadenze previste dovranno essere sottoposte a revisione e approvazione da parte della Stazione Appaltante, per garantire l'allineamento tra le aspettative delle parti coinvolte.

2.2.7. Garanzia

La garanzia fornita dall'aggiudicatario dovrà coprire un periodo di almeno 12 (dodici) mesi dalla data dal superamento della verifica di conformità della strumentazione, fatta salva l'offerta migliorativa presentata in sede di gara. Tale garanzia deve comprendere le riparazioni o sostituzioni di parti (con esclusione delle parti c.d. "consumabili" chiaramente individuabili nella documentazione a corredo) necessarie al funzionamento ottimale della strumentazione. Devono ritenersi, inoltre, comprese nella garanzia le spese di trasferta ed i costi della manodopera dei tecnici presso la sede di consegna ed installazione. Per l'intero periodo di vigenza della garanzia, l'aggiudicatario dovrà impegnarsi a fornire gratuitamente gli eventuali upgrade alle licenze software.

2.2.8. Assistenza tecnica, supporto e manutenzione

In caso di guasto l'aggiudicatario dovrà essere in grado di intervenire tempestivamente dalla segnalazione effettuata a mezzo PEC entro un massimo di 3 (tre) giorni lavorativi, fatta salva l'offerta migliorativa presentata in sede di gara. Tale intervento è finalizzato alla immediata assistenza ed al ripristino delle funzionalità dei software o, nel caso in cui ciò non sia possibile, alla valutazione delle criticità e degli interventi necessari.

3. MODALITÀ DI REALIZZAZIONE DEL SERVIZIO

Il servizio di progettazione della piattaforma software è un processo complesso che prevede le attività di analisi, progettazione, sviluppo, rilascio, consegna e installazione. Tutte le persone coinvolte nello svolgimento delle attività dovranno, quindi, operare in stretto coordinamento con lo staff preposto allo svolgimento delle attività FOSSR della sede CNR-ISTC di Roma;
La Stazione Appaltante dovrà essere costantemente informata sulle attività in corso in ogni fase, e dovrà essere coinvolta in tutte le scelte strategiche connesse alla buona riuscita del progetto. I risultati di ogni fase dovranno essere formalizzati in specifici documenti da trasmettere alla Stazione Appaltante.
Nelle prime fasi delle attività dovrà essere concordato il piano degli incontri periodici in cui valutare l'avanzamento dei lavori e le soluzioni proposte dalla Ditta Aggiudicataria.
È obbligatorio svolgere almeno 1 incontro ogni 2 settimane che potrà essere svolto anche in remoto attraverso l'utilizzo di strumenti quali Skype, MS Teams, Google Meet, ecc. Con cadenza mensile deve essere previsto obbligatoriamente un incontro face-to-face per il quale la Ditta Aggiudicataria dovrà garantire la presenza fisica di almeno due responsabili (o delegati) coinvolti nel progetto. Le riunioni dovranno essere pianificate con almeno 1 settimana di anticipo e il linguaggio utilizzato sarà l'italiano. Al termine della riunione, dovrà essere stilata apposita minuta sempre in italiano che la Stazione Appaltante e la Ditta Aggiudicataria dovranno approvare.
Tutte le attività di consulenza, inclusa l'installazione e la personalizzazione del software necessario, saranno condotte sui sistemi hardware di cui la SA dispone presso la sede di Roma di CNR-ISTC.
La Ditta Aggiudicataria dovrà fornire adeguata descrizione tecnica con le soluzioni progettuali ed implementative dettagliate utilizzando gli strumenti comuni dell'Ingegneria del Software quali UML, etc. o di Ingegneria della Conoscenza quali Protégé, Graffoo, VOWL, etc.
Nelle prime fasi delle attività dovrà essere:

- concordato il piano degli incontri periodici in cui valutare l'avanzamento dei lavori e le soluzioni proposte dalla Ditta Aggiudicataria;
- predisposto l'ambiente di condivisione di tutto il materiale riguardante la fornitura (es. codice sorgente, documenti di progettazione, verbali delle riunioni, schemi, ecc.) a cui la SA deve poter accedere in modo completo ed incondizionato.

È obbligatorio svolgere almeno 1 incontro ogni 2 settimane che potrà essere svolto anche in remoto attraverso l'utilizzo di strumenti quali Skype, Google Meet, ecc. Ogni 2 mesi deve essere previsto obbligatoriamente un incontro face-to-face per il quale la Ditta Aggiudicataria dovrà garantire la presenza fisica di almeno due responsabili (o delegati) coinvolti nel progetto. Le riunioni dovranno essere pianificate con almeno 1 settimana di anticipo e il linguaggio utilizzato potrà essere sarà l'italiano o l'inglese. Al termine della riunione, dovrà essere stilata apposita minuta in lingua inglese che la Stazione Appaltante e la Ditta Aggiudicataria dovranno approvare.

3.1. Piano di progetto

La durata complessiva del progetto è di **mesi 12** e comunque dovrà completarsi entro il **28.02.2025**. Il cronoprogramma di progetto da rispettare è riportato nella tabella successiva.

#Task	M1	M 2	M 3	M 4	M 5	M 6	M 7	M 8	M 9	M 10	M 11	M 12
Definizione degli aspetti operativi della fornitura	X											
Definizione dell'architettura della piattaforma	X	X										
Progettazione, Sviluppo, test e rilascio dei blocchi funzionali			X	X	X	X	X	X	X			
Integrazione dei blocchi funzionali, rilascio della piattaforma e popolamento dei dati								X	X	X	X	
Rilascio degli scenari operativi e collaudo										X	X	X
Formazione e servizi accessori												X

Rosario Falcone
07.12.2023
17:21:09
GMT+00:00