Scheda di Programma

Per l'attivazione nell'ambito del Corso di Dottorato di ricerca in MATEMATICA E SCIENZE COMPUTAZIONALI del seguente Programma di ricerca, a valere sulle risorse di cui al DM n. 351/2022, relativamente alla seguente Misura:

☐ **M4C1- Inv. 3.4** "*Didattica e competenze universitarie avanzate*" → **Dottorati dedicati alle transizioni digitali e ambientali.**

☒ **M4C1- Inv. 4.1** *"Estensione del numero di dottorati di ricerca e dottorati innovativi per la pubblica amministrazione e il patrimonio culturale".* In particolare:

☐ **Dottorati PNRR**

☒ **Dottorati per la Pubblica Amministrazione**
(selezionare l'area/le aree CUN di riferimento del programma tra quelle di seguito indicate)

- ☐ Area 09 – Ingegneria industriale e dell'informazione
- ☐ Area 11 – Scienze storiche, filosofiche, pedagogiche e psicologiche
- ☐ Area 12 – Scienze giuridiche
- ☐ Area 13 – Scienze economiche e statistiche
- ☐ Area 14 – Scienze politiche e sociali

☐ **Dottorati per il patrimonio culturale**
(selezionare l'area/le aree disciplinare/i e la tematica del programma tra quelle di seguito indicate)

- ☐ Area 01 – Scienze matematiche e informatiche **Tematica** – Informatica, patrimonio e beni culturali
- ☐ Area 02 – Scienze Fisiche **Tematica** – Fisica applicata al patrimonio culturale e ai beni culturali
- ☐ Area 03 – Scienze chimiche **Tematica** – Chimica, ambiente, patrimonio e beni culturali
- ☐ Area 04 Scienze della Terra **Tematica** – Georisorse minerarie per l'ambiente, il patrimonio e i beni culturali
- ☐ Area 05 Scienze Biologiche **Tematica** - Ecologia, patrimonio e beni culturali
- ☐ Area 08 – Ingegneria civile e Architettura **Tematiche** 1) Architettura, ambiente antropizzato, patrimonio e beni culturali 2) Architettura e paesaggio 3) storia dell'architettura; 4) Restauro; 5) Pianificazione e progettazione dell'ambiente antropizzato; 6) Design e progettazione tecnologica dell'architettura
- ☐ Area 10 Scienze dell'antichità, filologico-letterarie e storico -artistiche **Tematiche** 1) Archeologia; 2) Storia dell'arte; 3) Media, patrimonio e beni culturali
- ☐ Area 11 – Scienze storiche, filosofiche, pedagogiche, psicologiche **Tematiche** 1) Biblioteconoma; 2) Archivistica; 3) Storia del patrimonio e dei beni culturali 4) Paleografia; 5) Estetica; 6) Didattica dell'arte; 7) pedagogia dell'Arte
- ☐ Area 12 - Scienze giuridiche **Tematica** Diritto del patrimonio culturale
- ☐ Area 13 - Scienze Economiche e statistiche **Tematiche** 1) Economia della cultura e dell'arte 2) Economia e gestione delle imprese artistiche e culturali; 3) Statistica e Data Analytics per i beni culturali
- ☐ Area 14 Scienze Politiche e sociali **Tematiche** 1) Sociologia dei beni culturali 2) sociologia dell'ambiente e del territorio

Titolo del Programma di ricerca: Open Innovation per la competitività di Pubblica Amministrazione e PMI: Algoritmi su stringhe, Machine Learning e HPC per analisi e automazione massiva dei documenti gestionali.

Title of the Research Program: Open Innovation for the competitiveness of Public Administration and SMEs: Algorithms on strings, Machine Learning and HPC for analysis and massive automation of management documents.

❖ **Descrizione** (MAX 5000 CARATTERI SPAZI ESCLUSI)**:**

VERSIONE ITALIANA

La gestione di Big Data in ambito manageriale e amministrativo è di grande rilevanza per la Pubblica Amministrazione e le Imprese. Dati semi-strutturati come moduli e protocolli, fattura elettronica, fascicoli sanitari o schede documentali sono spesso generati, catalogati e conservati con standard disomogenei e tecnologie sub-ottimali, che impediscono l'implementazione di sistemi applicativi nonostante investimenti ingenti. Tecniche e protocolli informatici convenzionali sono spesso incompatibili con le necessità di post- processamento e interrogazione massiva. Ciò rende l'attività di information retrieval e system integration complessa e proibitiva dal punto di vista economico. Le mansioni amministrative e documentali spesso incidono in maniera decisiva sull'efficienza ed efficacia dei servizi offerti da PA e imprese, determinando talvolta l'impossibilità di garantire prestazioni e livelli assistenziali previsti per cittadini e aziende.

Il programma di ricerca si pone i seguenti obiettivi principali:

- migliorare significativamente framework, algoritmi, procedure e standard per il processamento e la gestione dei documenti disponibili per questi ambiti;
- rilasciare tecnologie, componenti e documentazioni Open Source che fungano da base di partenza accessibile e gratuita per i progetti di istituzioni pubbliche, PMI e Startup italiane.

Molti investimenti ICT della Pubblica Amministrazione sono destinati a progetti su piattaforme, dati e analytics e digitalizzazione dei processi. Tuttavia, questi progetti ampi e complessi necessitano di tecnologie di base, API e standard per la gestione documentale e l'interoperabilità. Proprio questi potrebbero trarre benefici dal percorso di dottorato proposto, consentendo la creazione di strumenti e framework riutilizzabili in tutti gli interventi ICT della PA.

Attraverso la creazione di librerie software Open Source, algoritmi di Open Data Science e standard applicativi, sarà possibile fornire strumenti pronti all'uso per i progetti di innovazione della PA. La diffusione di standard aperti potrà portare alla creazione di implementazioni omogenee e potenzialmente interoperabili *by design*. Inoltre, gli strumenti elaborati saranno maggiormente accessibili anche per Startup e PMI innovative in ambito ICT, con il risultato di rendere economico, veloce e aperto lo sviluppo di piattaforme e tecnologie innovative destinate alla PA.

La diffusione di conoscenze scientifiche e di librerie Open Source per la relativa implementazione potrà consentire un miglioramento diffuso dell'accessibilità di tecnologie moderne e performanti, in grado di impattare positivamente le prestazioni e il livello assistenziale offerto ai cittadini. Ad esempio, strumenti predittivi e di estrazione di informazione da basi documentali non strutturate su larga scala, consentirebbe di rendere più economici, tempestivi ed accessibili gli strumenti di gestione di pratiche documentali (es. cartelle cliniche) e ridurre i tempi di attesa dovuti ai processi amministrativi e documentali per i servizi alla persona, con evidenti ricadute sulle fasce della popolazione più deboli, che non hanno accesso a strutture private o di consulenza per snellire tali procedure.

Attraverso strumenti di estrazione di entità geografiche dai documenti (es. Indirizzi, coordinate, etc.), sarà possibile geo-referenziare informazioni e renderle integrabili su scala massiva per attività quali la gestione del demanio, delle risorse naturali e per il monitoraggio di attività ad alto impatto ambientale. La disponibilità gratuita di strumenti e tecniche di base per la creazione di software e piattaforme che integrino tali tecnologie consentirà di rendere più accessibili e diffondere tali interventi.

L'attività formativa e di ricerca verterà su:

- *Indicizzazione compressa di collezioni di dati altamente ripetitivi*: nei contesti considerati, collezioni massive di dati altamente ripetitivi necessitano di essere memorizzate, analizzate e interrogate. Le strutture dati necessarie per elaborazioni complesse richiedono ulteriore spazio di ordine di grandezza superiore. La presente attività di ricerca prevede di definire e utilizzare nuove strutture dati di indicizzazione compressa capaci di memorizzare tutte le informazioni necessarie alla ricerca e localizzazione efficiente di pattern in uno spazio vicino alla taglia dei dati compressi. I dati non necessiteranno pertanto di essere decompressi ma potranno essere analizzati direttamente in forma compressa.
- *Classificazione e confronto di dati strutturati e semi-strutturati*: studio e utilizzo di strumenti di natura combinatoria per l'estrazione di informazioni, per la classificazione e il confronto tra testi. Gran parte dei dati con cui si intende operare sono strutturati in modo gerarchico. Tipicamente ciò accade per esempio nell'ambito Open Data. Si intende definire pertanto nuove metodologie per il confronto e la classificazione di dati strutturati in modo gerarchico, con lo scopo di localizzare similarità sotto-strutturali dei documenti analizzati.
- *Survival analysis e time-series analysis*: si prevede di utilizzare modelli basati su automi a stati finiti e algoritmi su grafi, integrati con tecniche di machine learning, per supportare survival analysis e time series analysis.

ENGLISH VERSION

The management of Big Data in the managerial and administrative fields is of great importance for the Public Administration and Companies. Semi-structured data such as forms and protocols, electronic invoices, health records or document files are often generated, classified and stored with inconsistent standards and sub-optimal technologies, which prevent the implementation of application systems despite large investments. Conventional IT techniques and protocols are often incompatible with the need for post-processing and massive interrogation. This makes information retrieval and system integration activities complex and prohibitive from an economic point of view. Administrative and documentary tasks often have a decisive impact on the efficiency and effectiveness of the services offered by PA and businesses, sometimes resulting in the inability to guarantee services and levels of assistance provided for citizens and companies.

The research program has the following main goals:

- significantly improve frameworks, algorithms, procedures and standards for the processing and management of documents available for these areas;

- release Open Source technologies, components and documentation that serve as an accessible and free starting point for projects of public institutions, SMEs and Italian Startups.

Many ICT investments of the Public Administration are destined for projects on platforms, data and analytics and process digitization. However, these large and complex projects require core

technologies, APIs and standards for document management and interoperability. Precisely, these could benefit from the proposed doctoral program, allowing the creation of reusable tools and frameworks in all ICT interventions of the PA.

Through the creation of Open Source software libraries, Open Data Science algorithms and application standards, it will be possible to provide ready-to-use tools for PA innovation projects. The spread of open standards could lead to the creation of homogeneous and potentially interoperable implementations by design. Furthermore, the tools developed will also be more accessible for innovative startups and SMEs in the ICT field, with the result of making the development of innovative platforms and technologies for the Public Administration inexpensive, fast and open.

The dissemination of scientific knowledge and Open Source libraries for its implementation will allow for a widespread improvement in the accessibility of modern and high-performance technologies, capable of positively impacting the performance and level of assistance offered to citizens. For example, predictive tools and information extraction from unstructured document bases on a large scale, would make it possible to make the management tools for document practices (e.g. medical records) more economical, timely and accessible and reduce waiting times due to processes administrative and documentary services for personal services, with evident repercussions on the weakest sections of the population, who do not have access to private or consultancy structures to streamline such procedures.

By means of tools for extracting geographical entities from documents (e.g. Addresses, coordinates, etc.), it will be possible to geo-reference information and make it integrable on a massive scale for activities such as the management of state property, natural resources, and for monitoring activities such as high environmental impact. The free availability of basic tools and techniques for creating software and platforms that integrate these technologies will make it possible to make these interventions more accessible and disseminated.

Educational and research activity will concern:

- Compressed indexing of highly repetitive data collections: in the contexts considered, massive collections of highly repetitive data need to be stored, analyzed and interrogated. The data structures required for complex processing require additional space of a higher order of magnitude. The present research activity foresees the definition and use of new compressed indexing data structures capable of storing all the information necessary for the efficient search and localization of patterns in a space close to the size of the compressed data. The data will therefore not need to be decompressed but can be analyzed directly in compressed form.
- Classification and comparison of structured and semi-structured data: study and use of combinatorial tools for the extraction of information, for the classification and comparison between texts. Much of the data with which you intend to work is structured in a hierarchical way. Typically, this happens for example in the Open Data area. Therefore, we intend to define new methodologies for the comparison and classification of hierarchically structured data, with the aim of localizing sub-structural similarities of the analyzed documents.
- Survival analysis and time-series analysis: it is planned to use models based on finite state automata and algorithms on graphs, integrated with machine learning techniques, to support survival analysis and time series analysis.

❖ **PERIODO IN IMPRESA – CENTRI DI RICERCA – P.A.:**

Il Programma di ricerca sarà svolto in collaborazione con il seguente soggetto:

Ragione sociale: BuildNN Srl

Sede legale: Via Capri 44, 70022 Altamura (BA)

Rappresentante legale: Giacomo Barone

L'ente sopra citato ospiterà il dottorando beneficiario della borsa finanziata sulle risorse del DM 351/2022 per n. 12 mesi **(min 6 max 12)** nel corso del dottorato.

- **PERIODO ALL'ESTERO:**

Il Programma di ricerca prevede un periodo all'estero di n. 6 mesi **(min 6 max 18)** presso la seguente istituzione:

Université Gustave Eiffel (Parigi, Francia)

Si dichiara inoltre che il presente programma è conforme al principio "di non arrecare un danno significativo" (DHSH) ai sensi dell'art. 17 del regolamento (UE) 2020/852 in coerenza con gli orientamenti tecnici predisposti dalla Commissione Europea (Comunicazione della Commissione Europea 2021/C58/01) e garantisce il rispetto dei principi orizzontali del PNRR (contributo all'obiettivo climatico e digitale c.d. tagging, il principio della parità di genere e l'obbligo di protezione e valorizzazione dei giovani).