Borsa di studio attivata ai sensi di quanto disposto dal D.M. n. 1061 del 10/08/2021

Titolo del progetto: Mappe del patrimonio immateriale della cultura germanica e slava elaborate con l'ausilio di Machine Learning

La borsa sarà attivata sul seguente corso di dottorato accreditato per il XXXVII ciclo:
STUDI GERMANICI E SLAVI

Responsabile scientifico: Camilla Miglio
Area per la quale si presenta la richiesta: INNOVAZIONE
Numero di mensilità da svolgere in azienda: 6
Numero di mensilità da svolgere all'estero: 12 presso Charles University
Azienda: AND Srl Ambienti Narrativi Digitali – Via Zanella, 2A – 50123 Firenze
Il Dipartimento è disponibile a cofinanziare per un importo pari a euro: euro 10.000
Dipartimento finanziatore: DIPARTIMENTO DI STUDI EUROPEI, AMERICANI E INTERCULTURALI con delibera del 21.09.2021

Progetto di ricerca:
OBIETTIVI DELLA RICERCA

1. Costruire e interpretare mappe dinamiche visuali (concettuali, geoculturali, tematiche) o rappresentazioni reticolari per parole chiave di una vasta sezione del patrimonio culturale immateriale germanico e/o slavo, sulla base di corpora in grado di interoperare con altri patrimoni digitalizzati. Ciò porterà a risultati specifici nell'analisi delle fonti di un determinato corpus scritto; per esempio: il patrimonio tramandato dai fratelli Grimm nelle diverse edizioni a stampa e relativi manoscritti delle saghe, delle fiabe, e delle tradizioni scritte e orali ad esse connesse e testimoniate nei paratesti, nella corrispondenza e nelle carte private presenti in archivio e solo parzialmente digitalizzate; il patrimonio raccolto e solo parzialmente pubblicato da Božena Nemcová o da Karel Jaromír Erben nelle edizioni a stampa delle fiabe e dei materiali di carattere etnografico, testimoniato nelle corrispondenze private e in carte private presenti in archivio. Il modello di analisi ed elaborazione sviluppato si offre come una risorsa nuova per la ricerca, in quanto applicabile a tematiche diverse, e di innovazione per la didattica.
2. Sviluppare, in collaborazione con l'azienda di informatica umanistica partner e con la Facoltà di Matematica e fisica della Charles University di Praga, un software che fornisca risultati misurabili e riutilizzabili, come modello, in altre ricerche dello stesso tipo.

REALIZZAZIONE DELLA RICERCA
Il progetto di ricerca deve prevedere l'applicazione di tecnologie informatiche innovative (web semantico, sistemi interattivi, intelligenza artificiale) nell'ambito della gestione di corpora complessi digitalizzati (o da digitalizzare) appartenenti al patrimonio immateriale della cultura germanica e/o slava: testi e paratesti originali di saghe, fiabe, canzoni, filastrocche, racconti tradizionali, trascrizioni manoscritte o in formato audiovisivo di tradizioni orali di diversa natura e provenienza e di diversa datazione, dalle origini storiche delle rispettive lingue ai giorni nostri.
I dati da organizzare, correlare e processare possono essere: dati archivistici o supporti editoriali editi, metadati (dimensioni, materiali e stato di conservazione); testi inediti, anche manoscritti, oggetti parlanti, iscrizioni, canti, tradizioni orali, registrazioni; informazioni testimoniali sulle lingue e culture; luoghi reali e immaginari; parole chiave (nomi, luoghi, caratteristiche fisiche, epoche, ecc.), pattern; relazioni tra parole chiave/pattern, lingue e culture, in

senso sincronico e diacronico.

Si prevede che siano impiegate, accanto alle competenze specifiche del settore germanico o slavo di appartenenza, necessarie per l'individuazione delle ipotesi di lavoro e per la validazione dei risultati finali, anche le opportunità offerte da Data Analytics e Machine Learning (ML), anche con app educational e chatbot dedicate. Le tecnologie legate al web semantico permettono di analizzare quantità ingenti di dati, difficilmente correlabili con metodologie tradizionali non assistite dal computer; di visualizzare mappe dinamiche da sviluppare sull'asse diacronico e sincronico; di aggregare dati a partire da pattern dedotti da testi, paratesti, immagini, parole chiave scritte o audioregistrate.
Le tecnologie da impiegare comprendono non solo la digitalizzazione e catalogazione di testi e immagini, ma anche l'impiego del Natural Language Processing (NLP) sviluppato dalla Intelligenza Artificiale (AI), allo scopo di confermare o smentire ipotesi di ricerca; proporre nuove ipotesi non individuabili in base a uno studio tradizionale delle fonti, non assistito dalle tecnologie AI/ML.
Fra le caratteristiche del software, va considerata anche la possibilità di implementare una rappresentazione della conoscenza dove possano coesistere diverse strutture di dati. Competenze nell'ambito di Linked Data, Linked Open Data e Key Enable Technologies consentono il collegamento tra piattaforme diverse, portando all'elaborazione di un'infrastruttura contenente dataset innovativi, utili per mostrare e studiare le interrelazioni, nella lunga durata ed entro spazi geoculturali molto ampi, di tradizioni ed espressioni dell'immaginario che ciascuna cultura "nazionale" tende a considerare come "proprie".
L'applicazione di sistemi interattivi può riguardare anche la rappresentazione, condivisione e disseminazione dei risultati della ricerca avanzata, coinvolgendo la comunità educativa come stakeholder. Oltre che per raggiungere gli obiettivi primari indicati, i dati estratti e catalogati possono essere riorganizzati anche a fini di innovazione disciplinare (ricerca di base, didattica).
Presso la Charles University, università partner del dottorato internazionale congiunto in Studi germanici e slavi, il dottorando/a a cui sarà assegnata la borsa potrà lavorare per 12 mesi non solo presso la Facoltà di Lettere e filosofia e in archivi e biblioteche contenenti materiali oggetto della ricerca specifica, ma anche presso la Facoltà di Matematica e Fisica (in particolare con l'Institute of Formal and Applied Linguistics), dove matematici esperti di AI hanno elaborato specifici software utilizzati in ambito letterario (ad esempio, il loro THEaiTRobot ha generato un copione teatrale messo in scena in occasione del centenario di R.U.R. Rossum's Universal Robots di Karel apek); e inoltre in collaborazione con il gruppo di ricerca su questioni di versificazione (Versification research group, https://versologie.cz/v2/web_content/?lang=en) dell'Accademia delle Scienze della Repubblica Ceca.

FASE DI VALIDAZIONE
È prevista una fase di validazione del modello elaborato in base a quanto descritto: il modello proposto dovrà confermare un output in linea con quello della metodologia tradizionale.

FASE DI ITERAZIONE DEL MODELLO
A partire dai risultati ottenuti con Machine Learning, dopo opportuna validazione, si può avviare una fase successiva di ricerca di nuove ipotesi, impiegando lo stesso modello in maniera iterativa, allo scopo di individuare nuove correlazioni e risultati di ricerca più approfonditi.

Titolo del progetto (inglese): Maps from the intangible Germanic and Slavic cultural heritage processed through Machine Learning

Progetto di ricerca (inglese):
GOALS OF THE RESEARCH PROJECT

The research aims at:
1. Elaborating and interpreting dynamic visual maps (conceptual, geocultural, thematic) or keyword reticular representations of a large section of the Germanic and / or Slavic intangible cultural heritage, on the basis of corpora interoperating with other digitized heritages. This can lead to specific results in the analysis of the sources of a specific written corpus; a few examples: the heritage handed down by the Grimm brothers in the various printed editions and manuscripts of the sagas, fairy tales, and written and oral traditions connected to those and contained in the paratexts; the corpus collected and only partially published by Božena Nemcová or by Karel Jaromír Erben in the printed editions of fairy tales and ethnographic materials, in correspondence and private papers stored in archives and only partially digitized; contained in private correspondence and in private papers stored in the archives. The resulting analysis and processing model can represent a new resource for research, as it is applicable to different subjects, and an innovative tool for teaching.
2. Developing, in cooperation with the partner IT company and with the Faculty of Mathematics and Physics of Charles University in Prague, a software that is reusable in other researches in the cultural field.

REALIZATION OF THE RESEARCH PROJECT
The research project must include the use of innovative information technologies (semantic web, interactive systems, Artificial Intelligence) in the processing of complex digitized (or to be digitized) corpora belonging to the intangible heritage of Germanic and / or Slavic culture: texts and original paratexts of sagas, fairy tales, songs, nursery rhymes, traditional tales, handwritten or audiovisual transcriptions of oral traditions of different nature and origin and belonging to different times, from the historical origins of the respective languages to the present day.
The data to be organized, correlated and processed can be: archival data or published editorial supports, metadata (dimensions, materials and state of conservation); unpublished texts, including manuscripts, speaking objects, inscriptions, songs, oral traditions, recordings; information on the related languages and cultures; real and imaginary places; keywords (names, places, physical characteristics, eras, etc.), patterns; relationships between keywords/patterns, languages and cultures, in a synchronic and diachronic perspective.
We expect that, alongside the specific skills of the Germanic or Slavic sector to which they belong, necessary for imagining working hypotheses and for the validation of the final results, also the opportunities offered by Data Analytics and Machine Learning (ML) will be enhanced (for example through educational apps and dedicated chatbots). The technologies of the semantic web make it possible to analyze large amounts of data, which are difficult to correlate using traditional approaches, not assisted by the computer; view the dynamic maps to be developed on the diachronic and synchronic axis; aggregate data on the basis of patterns deduced from texts, paratexts, images, written or audio-recorded keywords.
The technologies to be used include not only the digitization and cataloging of texts and images, but also the use of Natural Language Processing (NLP) developed by Artificial Intelligence (AI) in order to: confirm or reject research hypotheses; propose new hypotheses that cannot be identified on the basis of a traditional study of sources that is not assisted by AI/ML technologies.

Among the features of the software, the possibility of implementing a representation of knowledge where different data structures can coexist must also be considered. Skills in Linked Data, Linked Open Data and Key Enable Technologies allow to connect different platforms, to develop an infrastructure containing innovative datasets, useful for showing and studying interrelationships, in the long term and in a wide geo-cultural perspective, between the traditions and the expressions that each "national" culture tends to consider as "its own" cultural heritage.
The application of interactive systems can also concern the representation, sharing and dissemination of the results of advanced research, involving the educational community as a stakeholder. In addition to achieving the above mentioned primary objectives, the extracted and catalogued data can also be reorganized for the purpose of innovation in the specific field of research (basic research, teaching).
At Charles University, partner university of the joint international doctorate in Germanic and Slavic Studies, the PhD student who will be awarded the scholarship will be able to work for 12 months not only at the Faculty of Arts and

Philosophy and in archives and libraries to find documents related to this specific research, but also at the Faculty of Mathematics and Physics (in particular in collaboration with the Institute of Formal and Applied Linguistics), where AI experts have developed specific softwares for the literary field (for example, their THEaiTRobot generated a play staged on the occasion of the centenary year of RUR Rossum's Universal Robots by Karel apek); and in collaboration with the Versification research group, https://versologie.cz/v2/web_content/?lang=en) of the Academy of Sciences of the Czech Republic.

VALIDATION

A validation of the model elaborated on the basis of the described results will be necessary: the proposed model has to confirm an output in line with the one provided by traditional methodologies.

ITERATION OF THE MODEL

Starting from the findings obtained through Machine Learning, after the above described validation phase, a following step can be activated to search for new hypotheses, using the said model iteratively, in order to identify new correlations and in-depth research findings.