"INVITO A PRESENTARE PROGETTI DI FORMAZIONE ALLA RICERCA IN ATTUAZIONE DEL PIANO TRIENNALE ALTE COMPETENZE PER LA RICERCA, IL TRASFERIMENTO TECNOLOGICO E L'IMPRENDITORIALITÀ APPROVATO CON DELIBERAZIONE DELL'ASSEMBLEA LEGISLATIVA N. 38 DEL 20/10/2015
POR FSE 2014/2020 Obiettivo tematico 10"

Delibera di Giunta regionale n. 388 del 19/03/2018
Scadenza 26/04/2018

scheda descrittiva dei Progetti di formazione alla ricerca

Titolo del progetto di formazione alla ricerca

Uso di tecnologie del Semantic Publishing per l'arricchimento semantico di pubblicazioni accademiche delle aree Umanistiche e delle Scienze Sociali

(selezionare uno dei seguenti ambiti)

☒ Ambito A) "RISORSE UMANE PER UN'ECONOMIA DIGITALE: BIG DATA"
☐ Ambito B) "RISORSE UMANE PER LA SPECIALIZZAZIONE INTELLIGENTE"
☐ Ambito C) "RISORSE UMANE PER IL PATRIMONIO CULTURALE"

Ateneo

Alma Mater Studiorum – Università di Bologna

Corso di dottorato e relativo il ciclo

Culture Letterarie e Filologiche – 34° ciclo

Coordinatore del Corso di dottorato

Prof. Luciano Formisano

Referente scientifico/supervisore borsa di dottorato:	*Telefono:*	*e-mail:*
Francesca Tomasi	+39 3473241920	francesca.tomasi@unibo.it
Co-referente:		
Silvio Peroni	+39 3487234548	silvio.peroni@unibo.it

Dipartimento
Dipartimento di Filologia Classica e Italianistica

Sintesi del progetto di formazione alla ricerca

L'editoria semantica (o Semantic Publishing) riguarda l'uso delle moderne tecnologie proprie del Web e del Web Semantico per l'arricchimento di pubblicazioni accademiche così da renderle più interattive e aperte, e in modo che siano facilmente scopribili mediante l'uso di software intelligenti capaci di semplificare l'attività di lettura e sintesi del dominio accademico. Negli ultimi anni, diversi investimenti di ricerca sono stati fatti per esplorare questo tema, basti pensare a OpenAIRE (https://www.openaire.eu) in ambito europeo, e a OpenCitations (http://opencitations.net) o a CATARSI (Bando AlmaIdea 2017) che vedono l'Università di Bologna direttamente coinvolta. Oltre a questo, l'industria editoriale internazionale ha altresì iniziato utilizzare queste tecnologie in modo sistematico, mettendo a disposizione del pubblico grandi quantità di dati (ad esempio, si veda SciGraph di Springer Nature, https://scigraph.springernature.com) o attivando progetti di ricerca su questi temi, come il progetto SCAR (http://dasplab.cs.unibo.it/index.php/semantic-coloring-academic-references/) frutto di una collaborazione tra il Dipartimento di Informatica dell'Università di Bologna e Elsevier. Tuttavia, le discipline accademiche delle pubblicazioni analizzate su cui questi progetti e collaborazioni si sono principalmente concentrati sono primariamente di caratterizzazione scientifica (Informatica, Chimica, Biologia, Medicina, etc.) e in lingua inglese.

Lo scopo del progetto di dottorato prevede lo studio di possibili strategie di sviluppo, riutilizzo ed di estensione di queste tecnologie in modo da poter prendere in considerazione l'analisi delle pubblicazioni e altri oggetti di ricerca di provenienza Umanistica e di quelle relative alle Scienze Sociali, discipline che non sono state mai analizzate con un dettaglio sufficientemente preciso dai progetti in corso. Probabilmente, questo è dovuto al fatto che le pubblicazioni e oggetti della ricerca relative a queste discipline sono tipicamente più lunghe (es. i libri), complesse (a livello argomentativo), e meno strutturate a priori, e quindi più difficili da analizzare rispetto a quelle tipicamente scientifiche. Inoltre, il progetto vuole altresì dare particolare rilevanza alle pubblicazioni in italiano, che finora non sono state oggetto di esaustive analisi mediante gli strumenti del Semantic Publishing, e quindi rappresentano un importante nuova copertura dell'applicabilità di queste tecnologie.

Gli obiettivi del progetto includono a) un'analisi ampia della struttura documentale e argomentativa delle pubblicazioni Umanistiche e delle Scienze Sociali in modo da poter addestrare appropriatamente sistemi automatici di classificazione e estrazione della semantica delle pubblicazioni espressa con formalismi logici e comprensibili dalle macchine, e b) il rilascio di queste informazioni all'interno di grosse collezioni di dati accademici aperti (o scholarly big data), come quelli messi a disposizione da OpenCitations e seguendo le linee guida promosse dall'Initiative for Open Citations (https://i4oc.org) e i FAIR principles (https://www.force11.org/group/fairgroup/fairprinciples), così da poterli mettere a disposizione della comunità per ulteriori analisi e applicazioni.

Considerando l'urgente richiesta attuale da parte dell'industria editoriale internazionale, formare degli specialisti regionali delle tecnologie del Semantic Publishing è, di conseguenza, un aspetto cruciale per far si che editori accademici presenti sul territorio possano reperire competenze tecniche sufficienti per adattare il loro processo editoriale e i loro servizi alle richieste di mercato attuali e future. L'obiettivo finale è quello di permettere agli editori di essere più di una semplice industria dedita alla pubblicazione di materiale accademico, orientandoli sempre di più verso la produzione di servizi utili alla comunità accademica.

Finalità generali e i risultati attesi del progetto di formazione alla ricerca

Il progetto di ricerca identifica due obiettivi principali, di cui uno teorico e uno applicativo.

Dal punto di vista teorico, si auspica uno studio approfondito dei modelli di descrizione semantica esistenti (ovvero, le ontologie) sviluppati per descrivere il dominio editoriale, come le SPAR Ontologies (http://www.sparontologies.net) e standard internazionali quali Dublin Core Metadata Terms (http://dublincore.org/documents/dcmi-terms/) e Functional Requirements for Bibliographic Records (https://www.ifla.org/publications/functional-requirements-for-bibliographic-records), in modo da poterli riadattare e/o estendere per coprire le esigenze delle discipline in considerazione, ovvero quelle Umanistiche e le Scienze Sociali.

A livello applicativo, invece, l'obiettivo è quello di mettere a disposizione degli strumenti software che permettano una più agevole integrazione delle suddette tecnologie all'interno di processi editoriali e di pubblicazione esistenti. Inoltre, si provvederà anche alla pubblicazione di questi dati e metadati nel pubblico dominio, all'interno di opportune collezioni di dati bibliografici come OpenCitations, che al momento non hanno a disposizione informazioni relative ai prodotti della ricerca delle discipline Umanistiche e delle Scienze Sociali. Il fine ultimo, in questo caso, è quello di mettere a disposizione questi dati in modo libero per permettere eventuali ulteriori analisi da parte di ricercatori esterni. In questo modo, il contributo di ricerca si inserisce perfettamente nelle odierne tematiche relative all'Open Science promosse dall'Unione Europea (https://ec.europa.eu/research/openscience/index.cfm) e dall'UNESCO (http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/open-science-movement/), e supportate da iniziative a caratura internazionale, come la Initiative for Open Citations (https://i4oc.org).

Le competenze acquisite dal candidato della borsa del dottorato alla fine del periodo di formazione permetteranno una immediata spendibilità a livello industriale, considerando la crescente esigenza da parte dell'editoria accademica di figure altamente competenti nell'uso delle tecnologie di Semantic Publishing. Inoltre, le stesse competenze possono altresì essere sufficienti anche per intraprendere un eventuale percorso accademico, nel caso in cui si voglia approfondire ulteriormente la componente di ricerca trattata dal progetto.

Inoltre, un ulteriore scopo della borsa di dottorato sarà ampliare l'analisi sul territorio regionale. Verranno coinvolte le case editrici della regione Emilia-Romagna al fine di verificare le possibilità di spendere gli strumenti realizzati per testare corpora bibliografici da esse pubblicati, in particolare le loro risorse bibliografiche già a disposizione in formato aperto, per proporre un nuovo modello di analisi dei dati. Sono già istituiti rapporti di collaborazione in particolare con le realtà attive sul territorio bolognese: Zanichelli, Il Mulino, BUP, Clueb.

Coerenza del progetto rispetto alla Strategia di specializzazione intelligente e alle linee programmatiche di sviluppo regionale

Il progetto di dottorato è allineato alle aspettative attese dalla Regione Emilia-Romagna relativamente al tema dei big data (https://formazionelavoro.regione.emilia-romagna.it/alta-formazione-ricerca/big-data) nel contesto delle Discipline Umanistiche e delle Scienze Sociali, e in particolare nell'ambito dell'Informatica Umanistica (o Digital Humanities). Infatti, il progetto si focalizza su uno specifico campo applicativo, ovvero quello della disponibilità dei dati e dei metadati relativi all'editoria accademica, che è riconosciuto essere una problematica di big data piuttosto affermata e di recente sviluppo – si veda a tal proposito l'articolo Xia et al. "Big Scholarly Data: A Survey", IEEE Transaction on Big Data, 3 (1), 2017, DOI: https://doi.org/10.1109/TBDATA.2016.2641460.

L'assenza di studi approfonditi dei modelli semantici di descrizione del contenuto dei prodotti della ricerca relativi alle discipline menzionate, così come il potenziale sviluppo di applicativi e interfacce che ne permettano la sintesi, l'analisi, l'identificazione di pattern ricorrenti e di reti di relazioni (reti di citazioni, reti di collaborazione tra autori e/o istituti di ricerca, etc.), permettono alla proposta di progetto di inserirsi nell'attuale sviluppo scientifico che è stato, finora, principalmente dedicato ai domini delle "scienze dure" come la Chimica e la Biologia. Inoltre, gli eventuali dati prodotti a seguito della ricerca, verranno contestualizzati e inseriti in ampie collezioni aperte di dati accademici (ad esempio OpenCitations, http://opencitations.net) e enciclopedici (ad esempio Wikidata, https://www.wikidata.org) in modo che possano essere liberamente riutilizzati dalla comunità locale, nazionale, e internazionale.

Per i progetti di cui all'ambito B) descrivere la focalizzazione rispetto alle value chain più rilevanti per l'economia regionale anche coerentemente con gli ambiti di attività dei Clust-ER

N/A

Conoscenze e competenze attese e descrizione della loro declinazione nel sistema economico produttivo

Le competenze create attraverso la formazione dottorale relativa al progetto integrano un approfondimento teorico nell'ambito dell'editoria accademica con l'acquisizione di competenze tecniche e tecnologiche proprie alla gestione dei dati in formato semantico mediante l'utilizzo dei moderni standard del Web e del Web Semantico.

Le case editrici della Regione Emilia-Romagna potranno trarre giovamento da un nuovo modello di analisi dei dati prodotti nel processo editoriale. Per questo la borsa di Dottorato formerà una figura di professionista capace di spendere le proprie competenze sul mercato editoriale, attraverso proposte innovative di valorizzazione della conoscenza, che nativamente risiede nelle pubblicazioni editoriali.

Spendibilità nel sistema economico produttivo delle conoscenze e competenze e analisi dei risultati occupazionali attesi

Le competenze sviluppate tramite la borsa di dottorato sono immediatamente spendibili nel contesto editoriale (ad esempio all'interno dagli editori accademici della regione Emilia-Romagna), in particolare nel momento in cui un eventuale editore voglia conformare i suoi processi di produzione a pratiche che attualmente vengono sperimentate e messe a frutto soltanto da grandi editori internazionali, quali, per esempio, Elsevier e Springer Nature.

In aggiunta, le competenze tecniche acquisite sulla creazione, manipolazione, analisi e visualizzazione di big data mediante le tecnologie del Web Semantico e del Semantic Publishing sono applicabili a vari domini di forte interesse regionale e nazionale, ad esempio per gestire e mettere a disposizione open data della pubblica amministrazione, catalogazioni museali, biblioteche Universitarie e comunali. Queste competenze sono, per esempio, già state richieste da diversi attori nello scenario regionale – basti ricordare la Fondazione Federico Zeri con il progetto Zeri&LODE (http://data.fondazionezeri.unibo.it) o l'Istituto per i Beni Artistici, Culturali e Naturali della regione con il progetto ReLOAD (https://labs.regesta.com/progettoReload/en).

Contestualizzazione del progetto: descrizione delle iniziative di ricerca e innovazione nelle quali si colloca la proposta

L'Università di Bologna è particolarmente attiva nel contesto dei big data, e vanta numerosi progetti che coinvolgono personale strutturato dell'Ateneo e che riguardano il dominio dell'editoria accademica e dell'umanistica digitale, e che utilizzano tecnologie del Web Semantico. Tra quelli attivi, si possono citare:

- OpenCitations (http://opencitations.net): una *service organisation* che mette a disposizione vari applicativi *open source* per la visualizzazione di dati semantici ed è responsabile per la creazione dell'OpenCitations Corpus (OCC), una collezione aperta di dati citationali di articoli accademici, e che vanta numerose collaborazioni con progetti e organizzazioni vicine al tema delle *open citations* – dipartimenti coinvolti: Dipartimento di Filologia Classica e Italianistica, e Dipartimento di Informatica - Scienza e Ingegneria;
- SCAR (http://dasplab.cs.unibo.it/index.php/semantic-coloring-academic-references/): progetto in collaborazione con Elsevier che ha come obiettivo la creazione e arricchimento della bibliografia presente negli articoli scientifici in modo da caratterizzarle secondo molteplici criteri, in particolare identificando la funzione delle varie citazioni (ovvero, il motivo per cui un articolo ne cita un altro) - dipartimenti coinvolti: Dipartimento di Informatica – Science e Ingegneria, con l'aggiunta di personale del Dipartimento di Filologia Classica e Italianistica;
- CATARSI: progetto finanziato dal Bando AlmaIdea 2017, prevede un'analisi degli strumenti automatizzati esistenti al fine dell'interpretazione automatica del contenuto informativo di un testo, articolo scientifico o di quotidiano - dipartimenti coinvolti: Dipartimento di Informatica - Scienza e Ingegneria, Dipartimento di Filologia Classica e Italianistica, Dipartimento di Scienze Aziendali;
- SPAR Ontologies (http://www.sparontologies.net): una collezione di modelli ontologici per la descrizione, in formati propri al Web Semantico, delle informazioni relative all'editoria, tra le quali la descrizione dei metadati documentali, gli identificativi delle risorse bibliografiche, le tipologie di citazioni e i relativi contesti citazionali, i riferimenti bibliografici, le parti che compongono un documento, i ruoli che persone e organizzazioni hanno all'interno del ciclo editoriale, i dati bibliometrici associati a risorse bibliografiche, e i processi editoriali - dipartimenti coinvolti: personale del Dipartimento di Filologia Classica e Italianistica;
- Zeri&LODE (http://data.fondazionezeri.unibo.it/): progetto di ricerca che concerne la pubblicazioni di parte consistente del catalogo online della Fototeca Zeri sotto forma di *Linked Open Data*, utilizzando i formati propri del Web Semantico – dipartimenti coinvolti: Dipartimento di Filologia Classica e Italianistica, Dipartimento di Informatica - Scienza e Ingegneria, e Fondazione Federico Zeri.

Contestualizzazione del progetto: descrizione dei progetti competitivi maggiormente rilevanti nei quali si colloca la proposta

Tra i vari progetti in corso che riguardano lo studio e la creazione di dati relativi a pubblicazioni delle Discipline Umanistiche e delle Scienze Sociali, si possono menzionare:

- EXCITE Project (http://west.uni-koblenz.de/en/research/excite/): è un progetto finanziato dalla *Deutsche Forschungsgemeinschaft* che prevede l'estrazione di citazioni da una collezione di documenti accademici del dominio delle Scienze Sociali – principali istituzioni coinvolte: University of Koblenz-Landau e GESIS;
- Linked Open Citation Database (LOC-DB, https://locdb.bib.uni-mannheim.de/blog/en/): è un progetto finanziato dalla *Deutsche Forschungsgemeinschaft* che prevede l'estrazione di citazioni da una collezione di documenti accademici del dominio delle Scienze Sociali – principali istituzioni coinvolte: Mannheim University Library;
- Linked Books (https://dhlab.epfl.ch/page-127959-en.html): è un progetto che si concentra nell'analizzare la relazione tra citazioni a documenti primari e quelle a sorgenti secondarie – Principali istituzioni coinvolte: École Polytechnique Fédérale de Lausanne.

Contestualizzazione del progetto: descrizione delle collaborazioni con soggetti pubblici e/o privati a livello nazionale e internazionale nelle quali si colloca la proposta

Al momento, non ci sono collaborazioni attive su questi temi con imprese e/o enti pubblici. Tuttavia, dato l'interesse della proposta nel mondo editoriale e accademico, esiste la possibilità di sfruttare rapporti formali e informali già stabiliti con editori e centri di ricerca all'interno della Laurea Internazionale di Secondo Livello "Digital Humanities and Digital Knowledge" dell'Università di Bologn, diretta dal responsabile scientifico della proposta di progetto.

Descrizione delle modalità con cui il beneficiario della borsa potrà essere coinvolto nelle iniziative/progetti/collaborazioni

Tutti gli istituti menzionati nelle precedenti sezioni (in particolare: University of Koblenz-Landau, GESIS, Mannheim University Library, Polytechnique Fédérale de Lausanne) sono soggetti naturali per attivare collaborazioni in merito al progetto, considerando la vicinanza dei temi trattati. Inoltre, i tre progetti menzionati in precedenza (EXCITE, LOC-DB, e Linked Books) hanno già attivato collaborazioni con OpenCitations (https://opencitations.wordpress.com/2018/03/23/early-adopters-of-the-opencitations-data-model/), che è gestito attivamente da personale del Dipartimento di Filologia Classica e Italianistica.

Descrizione delle eventuali collaborazioni con soggetti pubblici e/o privati che si intendono attivare per la realizzazione del progetto, da formalizzare prima dell'avvio del ciclo di dottorato di riferimento, indicando ruolo e contributo apportato, e accordi relativi alla proprietà intellettuale

La Società editrice il Mulino S.p.A. (https://www.mulino.it/) supporta esplicitamente i temi di ricerca proposti nel progetto, e si è impegnata a collaborare formalmente, fornendo materiale (ovvero i sorgenti) selezionato tra le loro pubblicazioni a carattere Umanistico e relative alle Scienze Sociali. A tal proposito, si veda la lettera di intenti in allegato, firmata dal legale responsabile.

Descrizione delle ricadute attese sul sistema regionale dell'innovazione e della ricerca nella prospettiva e nel contesto nazionale e internazionale

Oltre a creare una figura di rilievo capace di creare e gestire big data estratti a partire da pubblicazioni accademiche mediante l'utilizzo delle più moderne tecnologie del Web Semantico, tra i possibili risultati del progetto ci sarà anche la creazione e la messa a disposizione al pubblico – attraverso piattaforme già esistenti, come OpenCitations – di questi dati, in modo da andare ad arricchire la copertura delle informazioni ad oggi caratterizzata principalmente da dati provenienti dalle discipline scientifiche.