Relazione finale Assegno di ricerca

Assegnista di ricerca

Sara D'Abate

Titolo dell'assegno di ricerca

"ReAD - Representation of Architectural Data. Processi di conoscenza e valorizzazione del patrimonio architettonico mediante l'applicazione di tecnologie legate all'intelligenza artificiale" (CUP B55F21003420007)

Responsabile scientifico

Prof.ssa Paola Porretta, Università degli studi Roma Tre - Dipartimento di Architettura

Settore Scientifico Disciplinare (SSD) di riferimento

ICAR/19

Durata dell'assegno di ricerca

1 febbraio 2022 - 31 dicembre 2023 (1 annualità + 6 mesi di rinnovo + 5 mesi di proroga dovuta a astensione obbligatoria di maternità dal 07/06/2022 al 06/11/2022)

Descrizione delle attività svolte nell'ambito dell'assegno di ricerca

L'assegno di ricerca è stato svolto nell'ambito del progetto *ReAD | Representation of Architectural Data*, vincitore dell'avviso pubblico Progetti Gruppi di ricerca 2020 di Lazio Innova, POR FESR LAZIO 2014-2020, Asse I Ricerca e Innovazione (finanziamento: € 150.000), responsabili scientifici: Elisabetta Pallottino, Paola Porretta, per il Dipartimento di Architettura di Roma Tre (DArc-RM3); Aldo Gangemi, Valentina Presutti, per l'Istituto di Scienze e Tecnologie della Cognizione del Consiglio Nazionale per le Ricerche (ISTC-CNR); Carlo Birrozzi, Chiara Veninata, Fabrizio Magnani, per l'Istituto Centrale per il Catalogo e la Documentazione del Ministero della Cultura (ICCD-MiC).

Il progetto *ReAD | Representation of Architectural Data* ha sviluppato:

- una tecnologia computazionale per l'estrazione automatica di dati strutturati da fonti iconografiche e testuali non strutturate attraverso l'applicazione di metodologie di intelligenza artificiale di *image processing* e di *natural language processing*;

- il *knowledge graph* nel quale tali dati sono stati organizzati e resi accessibili agli utenti;
- la rete di ontologie tramite cui i dati sono stati semanticamente descritti e strutturati, che andrà ad ampliare *ArCO-Architettura della Conoscenza*, la più grande rete italiana di ontologie per la descrizione del patrimonio culturale sviluppata dagli Organismi di Ricerca proponenti del progetto ICCD-MiC e ISTC-CNR, con la collaborazione del DArc-RM3, nell'ambito della ricerca *Accesso e conservazione delle risorse digitali relative al patrimonio culturale* (2020-2021, responsabili scientifici: Elisabetta Pallottino, Michele Zampilli).

La ricerca si è articolata in differenti attività, strettamente correlate tra loro e volte al raggiungimento degli obiettivi intermedi e finali del progetto. Il lavoro svolto ha riguardato le attività relative ai seguenti Work Packages (WP) di progetto: *WP2 Definizione dei requisiti e individuazione delle fonti*, *WP3 Estrazione automatica di dati da fonti non strutturate*, *WP5 Validazione delle tecnologie e dei risultati*, *WP6 Diffusione, comunicazione e sfruttamento dei risultati*.

Tutte le attività sono state svolte all'interno del team di DArc-RM3 (con la supervisione delle responsabili del progetto prof.sse Elisabetta Pallottino e Paola Porretta), e in stretta collaborazione con ISTC-CNR e ICCD-MiC.

Nell'ambito di tali attività, ho coordinato e verificato il lavoro di 15 studenti di DArc-RM3 coinvolti nel progetto per lo svolgimento di "Altre attività formative". Gli studenti hanno preso parte a ReAD in qualità di esperti di dominio e hanno lavorato all'annotazione del dataset (vd. *ultra, T3.1 Creazione del dataset di addestramento*), all'analisi degli errori compiuti dal modello di *machine learning* (vd. *ultra, T5.2 Validazione della qualità dei dati estratti e correzione degli errori*) e alla validazione del lessico dei beni architettonici attraverso un tool appositamente sviluppato dal team di ISTC-CNR (vd. *ultra, T3.3 Realizzazione del modello NLP per l'estrazione di dati da fonti testuali*).

WP2 Definizione dei requisiti e individuazione delle fonti (febbraio 2022 - dicembre 2022)

Nelle attività del WP2 sono stati identificati i requisiti specifici dei prodotti finali di ReAD. Per definire tali requisiti, nell'ambito dell'assegno di ricerca, è stata condotta una ricognizione delle tecnologie esistenti di *machine learning* applicate al dominio dei beni architettonici. Tale indagine ha rilevato che, allo stato attuale, le ricerche si sono concentrate soprattutto sullo sviluppo di tecnologie di *image recognition* applicate a fotografie di architettura. Poco è stato fatto sui disegni di architettura e, in questi studi isolati, sono stati utilizzati quasi esclusivamente dataset composti da disegni digitali. L'applicazione di tecnologie di *image recognition* a disegni storici di architettura (e quindi non realizzati attraverso strumenti digitali) rappresenta un rilevante contributo innovativo.

In seguito alle rilevazioni dei bisogni degli stakeholder, condotte dal team afferente all'area disciplinare di Economia aziendale di Roma Tre (che collabora per il progetto con DArc-RM3), il lavoro è stato orientato alla definizione di:

1. lo *use case* della tecnologia, che ha indirizzato la prima fase dello sviluppo dei prodotti del progetto e che orienterà la loro futura implementazione. In qualità di Task Leader del Task *T2.4 Definizione dello use case*, ho coordinato le attività necessarie per l'elaborazione dello *use case*, che è stato modellato sulla base delle indagini condotte e delle conoscenze specifiche degli esperti del dominio dei beni architettonici del gruppo di ricerca, attraverso la definizione di un insieme di interazioni possibili tra il sistema e una serie di attori ad esso interessati. Lo *use case* è stato descritto attraverso la compilazione di "*user stories*", ovvero di storie scritte in linguaggio naturale che descrivono il possibile utilizzo del prodotto nel mondo reale, raccolte attraverso la piattaforma Github nell'apposito repository del progetto ReAD (https://github.com/read-project/stories). Le "*user stories*" mettono in relazione "Persone" (possibili utilizzatori dei prodotti ReAD) e "Scenari" (possibili utilizzi dei prodotti ReAD) con le relative "*Competency questions*" (cioè una serie di domande che identificano le richieste operative da sottoporre al software per il raggiungimento degli obiettivi). Per il momento, sono state individuate 9 "Persone" e 9 "Scenari". Queste potranno essere in futuro implementate, in modo da delineare ulteriori sviluppi delle tecnologie di ReAD. Lo *use case* è descritto nel deliverable *D2.4 Documento di descrizione dello use case* (vd. Sara D'Abate, Paola Porretta, Maria Chiara Frangipane, and Margherita Porena. D2.4 Documento di descrizione dello *use case*. Deliverable Progetto ReAD. 2022).

2. il corpus grafico e testuale di addestramento (*training set*) e di verifica (*test set*) della tecnologia computazionale di estrazione automatica di dati strutturati dal patrimonio documentale non strutturato relativo ai beni architettonici. Tale attività è stata svolta in stretta collaborazione con il team di ISTC-CNR e soprattutto con quello di ICCD-MiC, responsabile del Task *T2.3 Individuazione e reperimento delle fonti*. Infatti, per lo sviluppo delle tecnologie computazionali, ICCD-MiC ha reso disponibili le risorse digitali, grafiche (fotografie e disegni) e testuali, dell'Archivio fotografico del Gabinetto Fotografico Nazionale e del Catalogo generale dei Beni Culturali. Le immagini selezionate riguardano perlopiù il patrimonio architettonico pre-moderno italiano, sia civile che religioso, e in particolare i disegni sono realizzati con tecniche grafiche eterogenee e con diverse tipologie di rappresentazione. Il lavoro di selezione del corpus si è concluso con la redazione del Deliverable *D2.3 Elenco e descrizione delle fonti* (vd. Alessandro Coco, Sara D'Abate, Maria Chiara Frangipane and Paola Porretta. D2.3 Elenco e descrizione delle fonti. Deliverable Progetto ReAD. 2022).

3. le categorie di dati che la tecnologia computazionale ha dovuto estrarre automaticamente dalle fonti non strutturate. Queste sono state incluse nel *Documento di specifiche del progetto*, contenuto nel Deliverable *D2.2 Documento di specifiche*, la cui elaborazione è stata oggetto delle attività svolte nell'assegno di ricerca (vd. Sara D'Abate, Paola Porretta, Maria Chiara Frangipane, Margherita Porena, Fabio D'Amore. D2.2 Documento

di specifiche. Deliverable Progetto ReAD. 2022). Il documento, redatto in stretta collaborazione con il team di ISTC-CNR come sintesi finale delle attività del Task *T2.2 Individuazione delle categorie di dati significativi per la descrizione di beni architettonici*, contiene i requisiti tecnici e funzionali dei prodotti tecnologici che il progetto ha realizzato. In particolare, il documento descrive in maniera approfondita i requisiti funzionali che sono stati necessari per l'avvio della prima fase di sviluppo della tecnologia ed elenca più sommariamente le altre categorie di dati che potranno eventualmente essere oggetto dello sviluppo futuro del progetto, in caso di nuovi finanziamenti.

WP3 Estrazione automatica di dati da fonti non strutturate (novembre 2022 - agosto 2023)

Nell'ambito del WP3 mi sono occupata dei seguenti task:

- Task *T3.1 Creazione del dataset di addestramento*, in stretta collaborazione con il team di ISTC-CNR. Nell'ambito di questo task, ho svolto la prima fase di annotazione dei disegni architettonici dei dataset di riferimento, tramite l'impiego di Label Studio (un tool di classificazione delle immagini appositamente customizzato dall'assegnista di ricerca di ISTC-CNR). Successivamente, l'annotazione è stata ripetuta dagli studenti di DArc-RM3 coinvolti nel progetto. Scopo di questa attività è stato l'addestramento dell'algoritmo di *machine learning* per il riconoscimento della tipologia di disegno architettonico (pianta/prospetto/sezione).
Sono stati annotati due dataset: uno con i disegni estratti dal Catalogo generale dei Beni culturali tra i beni storico-artistici (3489 files) e uno con i disegni estratti dal sito Picryl (2924 files), ovvero un portale che raccoglie immagini con licenza in pubblico dominio. In particolare, per ogni immagine è stato annotato, in prima istanza, il tipo di proiezione ortogonale utilizzato (pianta/prospetto/sezione). In secondo luogo, sono state annotate, se rilevabili, alcune informazioni più specifiche relative al disegno: se rappresenta un dettaglio architettonico (es. una finestra, un capitello, una balaustra) o un oggetto a scala territoriale (es. un acquedotto, un parco, ecc) e se è realizzato a mano libera. L'obiettivo di queste annotazioni secondarie è stato quello di poter scremare il dataset in fase di test dell'algoritmo, al fine di ottenere un insieme quanto più possibile omogeneo di disegni tecnici che rappresentano edifici nella loro interezza alla scala architettonica e quindi di rendere più facile alla tecnologia il riconoscimento di caratteristiche simili.
In una eventuale futura implementazione del modello di *machine learning*, attuabile in caso di ulteriori finanziamenti del progetto, si potrà addestrare il tool al riconoscimento nei disegni della tipologia di edificio rappresentato (es. chiesa, anfiteatro, arco di trionfo ecc.). In quest'orizzonte, è stata già annotata la presenza di disegni raffiguranti chiese nei dataset di riferimento.
- Task *T3.3 Realizzazione del modello NLP per l'estrazione di dati da fonti testuali*. Ho eseguito la validazione di un vocabolario controllato per il dominio dei beni architettonici, al fine di costituire il "golden standard" sulla base del quale sono state valutate le validazioni compiute successivamente da altri esperti di dominio, ovvero dagli studenti del DArc-RM3. La validazione è stata effettuata attraverso un tool online appositamente sviluppato dai ricercatori dell'ISTC-CNR. Il lessico bilingue (italiano/inglese) è stato utilizzato sia nelle attività

legate allo sviluppo della tecnologia di *machine learning* (WP3), sia per la descrizione dei dati presenti nel *knowledge graph* (*WP4 Knowledge graph dei beni architettonici*), ed è stato reso accessibile agli utenti attraverso la pubblicazione sul repository Github del progetto.

WP5 Validazione delle tecnologie e dei risultati (marzo 2023-dicembre 2023)

Le attività si sono concentrate nell'ambito dei seguenti task:
- *T5.2 Validazione della qualità dei dati estratti e correzione degli errori*. La validazione della qualità dei dati estratti è stata effettuata attraverso l'analisi degli errori di predizione compiuti dal modello di *machine learning* in fase di test nella classificazione automatica di immagini volta al riconoscimento della tipologia di proiezione ortogonale utilizzata nella rappresentazione (pianta, prospetto, sezione). Tale attività, oltre ad essere utile a fini statistici, ha consentito di correggere alcuni errori (in particolare quelli legati a una annotazione non corretta delle immagini) e ha quindi permesso un progressivo miglioramento delle performance dei modelli di *machine learning*. Tale attività, proprio per il suo carattere iterativo, è stata ripetuta per due volte ed è stata svolta in collaborazione con i 15 studenti del Dipartimento di Architettura coinvolti nel progetto.
- *T5.3 Valutazione dei prodotti finali da parte degli stakeholders*. Per valutare la soddisfazione finale dei possibili utenti dei prodotti sviluppati da ReAD è stato sottoposto un questionario (appositamente sviluppato insieme con i ricercatori del CNR-ISTC) a 20 persone afferenti agli ambiti professionali correlati con il dominio di riferimento. Le rilevazioni hanno riguardato due prodotti: la *web application* per il riconoscimento automatico della tipologia di disegno architettonico (pianta, prospetto, sezione) sviluppata in seguito all'addestramento e verifica del modello di *machine learning* e il tool online di validazione del vocabolario controllato per il dominio dei beni architettonici. I dati ottenuti, successivamente analizzati e processati, hanno dimostrato un'accoglienza positiva dei prodotti ReAD da parte degli stakeholder.

WP6 Diffusione, comunicazione e sfruttamento dei risultati (maggio 2022-dicembre 2023)

Nell'ambito del Task *T6.2 Attivazione e mantenimento degli strumenti di comunicazione esterna e divulgazione*, sono state redatte circa 10 “Case history”, utili per descrivere gli obiettivi del progetto e per comunicare i suoi possibili utilizzi a un pubblico vasto, specializzato e non.

È inoltre stato organizzato un seminario dal titolo *Cultura digitale: prospettive di ricerca ed esperienze istituzionali. CNR-ISTC, MiC-ICCD, Università Roma Tre - Dipartimento di Architettura* a cura della scrivente e della prof.ssa Paola Porretta (in qualità di responsabile scientifica dell'assegno di ricerca), volto alla condivisione dei risultati raggiunti durante la ricerca con la comunità del Dipartimento. L'evento, programmato nell'ambito del Master biennale Internazionale di II livello *Culture del patrimonio. Conoscenza, tutela,*

valorizzazione, gestione dell'Università Roma Tre, modulo *Comunicazione e cultura digitale*, si è svolto martedì 11 luglio 2023, ore 10-13 e ha visto la partecipazione, in qualità di relatori, anche di Carlo Birrozzi (direttore ICCD-MiC), Fabrizio Magnani (funzionario ICCD-MiC), Margherita Porena e Maria Chiara Frangipane (ISTC-CNR).

Anche per la prossima edizione del Master (a.a. 2023-24) è prevista l'organizzazione di una giornata in cui saranno presentati i risultati finali del progetto di ricerca.

Nell'ambito del Task *T6.4 Attività di diffusione dei risultati*, ho presentato l'esito del lavoro in occasione di due convegni:

- XIV Convegno internazionale AIES *Diagnosi, conservazione e valorizzazione del patrimonio culturale*, Napoli, 14-15 dicembre 2023, organizzato da "AIES BBCC - Associazione Italiana di Esperti Scientifici Beni Culturali", con il paper dal titolo *ReAD: Representation of Architectural Data. Enhancement of Architectural Heritage through the Application of Artificial Intelligence*, insieme con C. Birrozzi, A. Coco, F. D'Amore, M.C. Frangipane, A. Gangemi, F. Magnani, E. Pallottino, M. Porena, P. Porretta, V. Presutti, C. Veninata (paper soggetto a procedura di selezione). Il paper è in corso di pubblicazione negli atti del convegno.

- Giornata di studio *Dialogare con l'imperdibile. Ontologie e grafi di conoscenza per la governance dei dati sul patrimonio culturale*, Bologna, 6 dicembre 2023, organizzato da Alma Mater Studiorum - Università di Bologna e ICCD, con la relazione *Il progetto ReAD -Representation of Architectural Data*, insieme con E. Pallottino, P. Porretta (su invito).

Roma, 22/12/2023

L'assegnista di ricerca

Il responsabile scientifico