CAD (Codificare, Analizzare, Diffondere) è una scuola di digital humanities finanziata dall'Associazione Internazionale Professori di Italiano (AIPI) con il contributo dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD).

Codificare, Analizzare, Diffondere: Le *Digital Humanities* nei progetti di ricerca

Programma

Il programma si articola su quattro giorni e prevede l'inizio delle attività il **martedì alle 10:30** e la fine il **venerdì alle 17:00**.

Si prevedono due tipi di attività: i **corsi** e i **laboratori** :

- Durante i **corsi** saranno presentati strumenti e tecnologie, tramite lezioni frontali e esercizi pratici. I primi due giorni i partecipanti seguiranno i corsi *Dalla carta allo schermo* e *Codifica dei testi*. Gli ultimi due giorni i partecipanti saranno divisi in due gruppi e seguiranno *Analisi semantica e di stile* oppure *Trattamento automatico del linguaggio*; e *Cartografia e network* oppure *Diffusione della ricerca online*. Una descrizione dettagliata dei corsi è disponibile qui di seguito.
- I **laboratori** prevedono l'approfondimento di quanto imparato durante i corsi e il lavoro pratico sui progetti di ricerca.

Le prime due giornate si chiudono con uno **sportello aperto**, individuale e facoltativo dove i partecipanti potranno porre delle domande dirette ai formatori, al fine di risolvere dubbi riguardanti le attività della giornata o direttamente sui casi di studio della loro ricerca.

Martedì 16 luglio

Ore	Sessione
10:15-11:00	Benvenuto
11:00-12:00	"Aperitivo" Showcase dei corsi
12:00-13:00	Presentazione progetti di ricerca
13:00-14:30	Pranzo
14:30-15:30	Presentazione progetti di ricerca

Ore	Sessione
15:30-16:00	Pausa caffé
16:00-17:00	CORSO: Dalla carta allo schermo
17:00-17:15	Divisione in gruppi
17:15-18:15	Keynote: Prof. Elena Pierazzo (Université de Tours)
18:15-19:00	Controllo computer portatili e sportello individuale (facoltativo)
19:00-…	Aperitivo

Mercoledì 17 luglio

Ore	Sessione
9:30-11:00	CORSO: Codifica dei testi
11:00-11:30	Pausa caffé
11:30-13:00	CORSO: Codifica dei testi
13:00-14:30	Pranzo
14:30-16:00	LABORATORI
16:00-16:30	Pausa caffé
16:30-18:00	LABORATORI
18:00-19:00	Sportello individuale (facoltativo)

Giovedì 18 luglio

Ore	Sessione
9:30-11:00	CORSO: Trattamento Automatico del Linguaggio / Analisi semantica e di stile
11:00-11:30	Pausa caffé

Ore	Sessione
11:30-13:00	CORSO: Trattamento Automatico del Linguaggio / Analisi semantica e di stile
13:00-14:30	Pranzo
14:30-16:00	LABORATORI
16:00-16:30	Pausa caffé
16:30-17:30	LABORATORI
17:30-18:30	Keynote: Prof. Fabio Ciotti (Università di Roma Tor Vergata)
20:00-…	Cena sociale

Venerdì 19 luglio

Ore	Sessione
9:30-11:00	CORSO: Cartografia e network / Diffondere la ricerca online
11:00-11:30	Pausa caffé
11:30-13:00	CORSO: Cartografia e network / Diffondere la ricerca online
13:00-14:30	Pranzo
14:30-16:00	LABORATORI
16:00-16:30	Pausa caffé
16:30-17:00	Feedback e saluti

I corsi nel dettaglio

Dalla carta allo schermo (Stefano Bazzaco)

Il corso offre uno sguardo sulle tecnologie OCR attualmente a disposizione che consentono la trasformazione di fotoriproduzioni di opere cartacee, manoscritte o a stampa, in testi processabili da parte del computer. Nello

specifico, il corso si divide in tre sezioni. In un primo momento, saranno presi in considerazione i recenti sviluppi nel campo delle tecnologie OCR e saranno presentati alcuni software OCR, proprietari o di libero accesso, di grande utilità e supporto per il lavoro degli umanisti (FineReader, Transkribus, OCRopus/Ocropy, etc.). In seguito, si presenteranno alcuni casi di studio e applicazione delle tecnologie di riconoscimento automatico e semi-automatico dei caratteri, prestando speciale attenzione all'ambito della trascrizione di fonti antiche e manoscritte, un'area di studio tuttavia in sviluppo in cui non sempre i sistemi OCR assicurano risultati ottimali. Infine, si forniranno indicazioni circa l'utilizzo delle piattaforme presentate e si chiariranno gli eventuali dubbi dei partecipanti in relazione ai loro specifici percorsi di ricerca.

Stefano è assegnista di ricerca presso il Dipartimento di Letterature, Lingue e Linguistica dell'Università di Verona. Durante i **laboratori** con lui potrete lavorare su OCR, sulla codifica dei testi in TEI, sulla stilometria.

Codifica dei testi (Simon Gabay)

Il corso propone un'introduzione alla codifica TEI (Text Encoding Initiative), uno dei più importanti standard del mondo dell'informatica umanistica (Digital Humanities). Durante il corso, i partecipanti impareranno a trasformare un testo in una base dati, interrogabile e pubblicabile in diversi formati (LaTeX, HTML, etc.). La scoperta dell'XML-TEI sarà anche l'occasione di conoscere ed adottare le norme per lavorare con il digitale, necessarie per rendere i propri dati riutilizzabili da altri ricercatori, interoperabili con altri sistemi e persistenti sul lungo termine. Dopo un'introduzione teorica ai principi dell'XML e al vocabolario della TEI, passeremo a degli esercizi che toccano i principali problemi sollevati dalla codifica, come la struttura fisica e logica del documento (pagine, titolo, capitoli, paragrafi, etc.) o le entità nominali (nomi di persone e di luoghi). Gli esercizi saranno l'occasione di presentare alcuni degli strumenti disponibili per facilitare il lavoro con la TEI, come il programma oXygen e l'applicazione web Roma.

Simon è ricercatore post-doc e insegna le Digital Humanities all'Università di Neuchâtel. Durante i **laboratori** con lui potrete lavorare su codifica di forme testuali specifiche (corrispondenze, teatro, etc.) e di casi filologici complessi (testimoni multipli, annotazione linguistica), integrazione di TEI con il trattamento automatico del linguaggio, OCR, stilometria e cartografia.

Analisi semantica e di stile (Simone Rebora)

Il corso si concentra su due aree principali dell'analisi del testo con metodi computazionali. Sotto il nome di "stilometria", si raccolgono una serie di approcci che hanno come fine ultimo quello di distinguere e misurare lo stile autoriale. Metodi statistici come la "Delta distance" vengono frequentemente utilizzati per l'attribuzione di testi anonimi, mentre la "keyness analysis" è adottata per individuare le marche lessicali che caratterizzano la scrittura di uno o più autori. Dopo aver passato in rassegna sinteticamente i fondamenti teorico-matematici per ognuno di questi metodi, verranno mostrate le molteplici modalità di visualizzazione dei risultati (come i dendrogrammi e gli alberi di consenso), che rendono infine possibile una "lettura da lontano" (distant reading) dei più vasti corpora testuali. Nella più estesa area dell'analisi semantica, a seconda delle esigenze dei partecipanti, verranno presentati: algoritmi di "sentiment analysis", che quantificano gli aspetti emotivi del testo con il fine di visualizzarne la struttura narrativa; algoritmi per la classificazione delle aree semantiche, che misurano le dominanti tematiche del testo; algoritmi di "topic modeling" e di semantica distribuzionale, che estraggono i temi e le relazioni concettuali direttamente dalla distribuzione delle parole in ampie collezioni testuali. Di tutti questi approcci saranno mostrate alcune semplici applicazioni negli studi letterari, sottolineando anche rischi e problematicità nel loro uso indiscriminato.

Simone è ricercatore post-doc all'Università di Basel e insegna letteratura comparata all'Università di Verona. Durante i **laboratori** con lui potrete lavorare su stilometria (stylo, JGAAP, pyDelta, pyZeta e quanteda), sentiment analysis (syuzhet, Stanford SA), classificazione semantica (LIWC, SEANCE), topic modeling (Gensim, LDA, Mallet) e semantica distribuzionale (word2vec, doc2vec).

Trattamento Automatico del Linguaggio (Greta Franzini)

Il corso è diviso in due parti: la prima parte si concentra sull'uso della linea di comando (conosciuta anche come "Command Line" o "Terminal") per la gestione di file di sistema e per la manipolazione di testi o corpora (i.e., pulizia e formattazione con Regular Expressions o RegEx), mentre la seconda è volta all'apprendimento dello strumento TreeTagger per il PoS-tagging (assegnazione automatica di parti del discorso) e la lemmatizzazione di testi

in lingua italiana. Una conoscenza, seppur elementare, della linea di comando è essenziale per tutti coloro che desiderano processare dati testuali automaticamente con strumenti NLP non muniti di interfacce grafiche sia su macchine portatili che su server (per analisi computazionali a larga scala). Egualmente essenziali sono il PoS-tagging e la lemmatizzazione quali task necessari per qualsiasi tipo di analisi linguistica computazionale, sia essa sintattica o semantica. Come per gli altri corsi di CAD, il corso TAL è diviso in parti uguali tra presentazioni frontali e esercizi pratici, accompagnati da istruzioni dettagliate. Alla fine del corso i partecipanti saranno non solo in grado di formattare e preparare testi o corpora in lingua italiana per vari tipi di analisi computazionali ma avranno anche acquisito le conoscenze necessarie per approfondire ed implementare quanto appreso nel loro attuale (o futuro) progetto di ricerca, consapevoli dei limiti e delle problematiche dei metodi computazionali in questo settore.

Greta è ricercatrice post-doc nel progetto ERC *LiLa: Linking Latin* presso l'Università Cattolica del Sacro Cuore (Milano). Durante i laboratori, con lei potrete lavorare su: linea di comando (e.g., pulizia/manipolazione di testi con Regular Expressions), trattamento automatico del linguaggio (PoS-tagging e lemmatizzazione in particolare), base dati relazionali (e.g., MySQL), versioning con GitHub/GitLab, pubblicazione web (HTML, CSS, Bootstrap) e codifica dei testi in TEI.

Cartografia e network (Giovanni Pietro Vitali)

Il corso di Cartografia e network è diviso in due moduli di un'ora e mezza ciascuno. L'obiettivo del corso è quello di far acquisire ai partecipanti gli strumenti tecnici basici per la visualizzazione di dati georeferenziati su mappe online e di dati relazionabili all'interno di network personalizzabili.

1. Gephi. Nella prima parte verrà affrontato l'utilizzo di un software chiamato Gephi, uno strumento fondamentale per la creazione di network, comunemente diffuso e utilizzato da tutta la comunità scientifica. Gephi verrà presentato attraverso degli esempi pratici "hands on" tali da mostrare quelle che sono le potenzialità di questo software. Saranno mostrati dei casi operativi all'interno dei quali risulti necessario fare ricorso a dei plugins, strumentali ad un diverso tipo di interazione coi dati. La necessità di sviluppare diversi tipi di grafici relazionali, verrà mostrata attraverso l'integrazione di Gephi con altri

strumenti, specialmente quelli legati alla cartografia digitale, al centro della seconda parte del presente corso.

2. Cartografia e coordinate. L'utilizzo di Gephi prevede una formattazione specifica per il dataset utilizzato. Tale adeguamento dei dati alle richieste di Gephi può essere facilmente commutato in un formato adatto anche alla realizzazione di carte digitali. In questa seconda parte del corso saranno mostrate le modalità di formattazione dei dati per la creazione di mappe con un'attenzione particolare all'estrazione delle coordinate sulla base di una lista di toponimi. In seguito verrà affrontata la creazione delle stesse mappe, a punti e a poligoni, attraverso delle applicazioni in linea quali Carto e Recogito, due strumenti semplici da usare per pubblicare carte digitali in rete.

Al termine del corso ogni partecipante sarà in grado di mettere in rete una carta con i propri dati per diffondere le sue ricerche o verificare le sue teorie.

Giovanni si occupa di lingua e letteratura contemporanea ed è Marie Skłodowska Curie Fellow a University College Cork. Durante i **laboratori** con lui potrete lavorare sul mapping, l'analisi dei network, la stilometria e il trattamento automatico del linguaggio.

Diffondere la ricerca online (Elena Spadini)

Il corso si concentra sulle tecnologie del web, in particolare la codifica HTML e i fogli di stile CSS. Avere familiarità con questi linguaggi permette di creare un sito web, ma anche di gestire la pubblicazione di dati in altri linguaggi (ad esempio, in TEI) e di controllare al meglio la pubblicazione tramite piattaforme come Wordpress o Omeka. Oltre alla creazione di siti e blog scientifici, si passeranno in rassegna altre modalità di diffusione della ricerca e di collaborazione online: i depositi istituzionali e internazionali (ad esempio, Zenodo), l'utilizzo di identificanti perenni (ad esempio, DOI), i software per il versioning (git) e le piattaforme per la gestione del code source (come Github e Gitlab). Le questioni legate all'Open Science (tra cui l'Open Access) e alle licenze per dati e pubblicazioni saranno brevemente trattate. Il corso è diviso in parti uguali tra presentazioni frontali e esercizi pratici, accompagnati da istruzioni dettagliate. Gli esercizi prevedono un livello base, per chi si avvicina a questi strumenti per la prima volta, e delle possibilità di approfondimento, per chi è già a suo agio. Alla fine del corso i partecipanti saranno in grado di creare un sito web statico per presentare il proprio progetto o il proprio

profilo professionale; e avranno acquisito le conoscenze necessarie per definire una strategia di gestione e diffusione dei dati e dei risultati della ricerca sul web, da implementare nell'attuale progetto o inserire nella domanda per un futuro progetto di ricerca.

Elena è ricercatrice post-doc all'Università di Losanna. Durante i **laboratori** con lei potrete lavorare su filologia digitale (collazione automatica, strumenti per la filologia romanza), modellazione dei dati e web semantico (base-dati relazionali, ontologie, OWL, RDF), XML (TEI, XSLT, pubblicazione di testi in TEI), pubblicazione web (HTML, CSS, Bootstrap).
