europass

Marta Soricetti

Nazionalità: Italiana **Data di nascita:** 29/12/1999

Numero di telefono: (+39) 3451254929

Indirizzo e-mail: marta.soricetti9@gmail.com

Abitazione: Via Aldo Moro 18B, 62019 Recanati (Italia)

ESPERIENZA LAVORATIVA

Assegnista di ricerca

Alma Mater Studiorum Università di Bologna [01/04/2024 – Attuale]

Città: Bologna | **Paese:** Italia

Il progetto di ricerca di cui mi occupo prevede lo sviluppo di CEC (https://github.com/opencitations/cec.git), un software per l'annotazione automatica di citazioni nel corpo del testo di articoli accademici forniti in formato PDF.

- sviluppo software
- sviluppo applicazioni web e REST APIs
- creazione di corpus per training e evaluation di strumenti di machine learning (Grobid, https://github.com/kermitt2/grobid)

Analista programmatrice

Stesi Consulting [13/11/2023 – 29/01/2024]

Paese: Italia

- supporto nell'analisi e interpretazione delle richieste delle aziende clienti
- personalizzazione ERP Odoo
- sviluppo nuovi moduli e applicazioni Odoo in Python
- utilizzo di strumenti di versionamento Git e GitHub

Tirocinio curriculare (12 cfu) - Attività di ricerca presso il DH.ARC (Digital Humanities Advanced Research Centre)

Alma Mater Studiorum Università di Bologna [16/10/2023 – 22/12/2023]

Città: Bologna | **Paese:** Italia

Ho svolto il tirocinio curriculare da 12 cfu previsto dal corso di studi in Digital Humanities and Digital Knowledge presso il centro di ricerca DH.ARC, e più nello specifico nel contesto di OpenCitations, per un totale di 350 ore.
Le principali attività di cui mi sono occupata sono le seguenti:

- Sviluppo Software, Sviluppo Test e Adattamento di Codice Preesistente in seguito a cambiamenti nel processo di ingestione dati dell'infrastruttura. In particolare, ho adattato il codice che era stato utilizzato in precedenza per l'ingestione di dati citazionali e bibliografici da DataCite al nuovo workflow che viene attualmente utilizzato in OpenCitations per l'ingestione di tutte le nuove sorgenti dati. Linguaggi di programmazione utilizzati: Python;
- Analisi e Visualizzazione Dati;
- Scrittura articolo accademico.

Tirocinio curriculare (6 cfu) - Attività di ricerca presso il DH.ARC (Digital Humanities Advanced Research Centre)

Alma Mater Studiorum Università di Bologna [15/06/2023 – 19/07/2023]

Città: Bologna | **Paese:** Italia

Ho svolto il tirocinio curriculare da 6 cfu previsto dal corso di studi in in Digital Humanities and Digital Knowledge presso il centro di ricerca DH.ARC, e più nello specifico nel contesto di OpenCitations, per un totale di 150 ore.

Le principali attività di cui mi sono occupata sono le seguenti:

- Sviluppo Software e Sviluppo Test per adattamento di codice preesistente in seguito a cambiamenti nel processo di ingestione dati dell'infrastruttura. L'adattamento in questo caso specifico ha interessato il codice per l'ingestione di dati da Crossref e non è stato sviluppato da me in prima persona. Io mi sono occupata della fase di testing e debugging, con la conseguente implementazione delle modifiche necessarie. Linguaggi di programmazione utilizzati: Python (unittest library);
- Analisi e Visualizzazione Dati.

ISTRUZIONE E FORMAZIONE

Laurea magistrale in Digital Humanities and Digital Knowledge

Alma Mater Studiorum Università di Bologna [01/08/2021 – 11/03/2024]

Città: Bologna | **Paese:** Italia | **Sito web:** https://corsi.unibo.it/2cycle/DigitalHumanitiesKnowledge | **Voto finale:** 110 con lode | **Tesi:** Ingesting JaLC Citation Data into OpenCitations: Methodology and Implementation

Esami sostenuti:

- primo anno: Computational management of data, Text retrieval, analysis and mining, Scholarly editing and digital approaches, Knowledge management, Information modeling and web technologies, Digital text in the humanities: theories, methodologies and applications, Digital heritage and multimedia;
- secondo anno: Network analysis, Natural language processing, International and eu copyright law, Business strategy and innovation in cultural industries, Open science, Basic analytics

Laurea triennale in Lettere (curriculum moderno)

Alma Mater Studiorum Univeristà di Bologna [01/08/2018 – 16/07/2021]

Città: Bologna | **Paese:** Italia | **Sito web:** https://corsi.unibo.it/laurea/lettere | **Voto finale:** 105 | **Tesi:** Ancona nel XII secolo e l'assedio del 1173

Esami sostenuti:

- primo anno: Informatica di base, Idoneità lingua inglese - b1, Laboratorio di lingua italiana, Letteratura italiana, Lingua latina, Linguistica generale, Storia romana;
- secondo anno: Letteratura italiana contemporanea, Filologia della lettaratura italiana, Letteratura latina, Storia dell'arte contemporanea, Linguistica italiana;
- terzo anno: Antropologia culturale, Geografia, Informatica umanistica, Letterature comparate, Storia medievale

Diploma di Liceo scientifico

Liceo Giacomo Leopardi Recanati [2013 – 2018]

Città: Recanati | **Paese:** Italia | **Sito web:** https://liceorecanati.edu.it/ | **Voto finale:** 100

COMPETENZE LINGUISTICHE

Lingua madre: italiano

Altre lingue:

inglese

ASCOLTO B2 **LETTURA** B2 **SCRITTURA** B2

PRODUZIONE ORALE B2 **INTERAZIONE ORALE** B2

Livelli: A1 e A2: Livello elementare B1 e B2: Livello intermedio C1 e C2: Livello avanzato

COMPETENZE DIGITALI

Programmazione

Conoscenza avanzata di Python e delle principali librerie (NumPy, Pandas, scikit-learn, Matplotlib) / Web Semantico: RDF, SPARQL, OWL / HTML, CSS, JAVASCRIPT / Realizzazione e utilizzo database con MySQL / Versionamento: Git, GitHub

Analisi e Visualizzazione Dati

Data Analysis Python / Microsoft Excel

PUBBLICAZIONI

[2024]

The OpenCitations Index: description of a database providing open citation data **Abstract**: This article presents the OpenCitations Index, a collection of open citation data maintained by OpenCitations, an independent, not-for-profit infrastructure organisation for open scholarship dedicated to publishing open bibliographic and citation data using Semantic Web and Linked Open Data technologies. The collection involves citation data harvested from multiple sources. To address the possibility of different sources providing citation data for bibliographic entities represented with different identifiers, therefore potentially representing same citation, a deduplication mechanism has been implemented. This ensures that citations integrated into OpenCitations Index are accurately identified uniquely, even when different identifiers are used. This mechanism follows a specific workflow, which encompasses a preprocessing of the original source data, a management of the provided bibliographic metadata, and the generation of new citation data to be integrated into the OpenCitations Index. The process relies on another data collection—OpenCitations Meta, and on the use of a new globally persistent identifier, namely OMID (OpenCitations Meta Identifier). As of July 2024, OpenCitations Index stores over 2 billion unique citation links, harvest from Crossref, the National Institute of Heath Open Citation Collection (NIH-OCC), DataCite, OpenAIRE, and the Japan Link Center (JaLC). OpenCitations Index can be systematically accessed and queried through several services, including SPARQL endpoint, REST APIs, and web interfaces. Additionally, dataset dumps are available for free download and reuse (under CC0 waiver) in various formats (CSV, N-Triples, and Scholix), including provenance and change tracking information.

Heibi, I., Moretti, A., Peroni, S., Soricetti M. 2024. Scientometrics

[2024]

The Integration of the Japan Link Center's Bibliographic Data into OpenCitations: The production of bibliographic and citation data structured according to the OpenCitations Data Model, originating from an Anglo-Japanese dataset **Abstract**: In this article, we present OpenCitations' main data collections: the unified index of citation data (OpenCitations Index), and the bibliographic data corpus (OpenCitations Meta) in view of the integration of a new dataset provided by the Japan Link Center (JaLC). Based on a computational analysis of the titles of the publications performed in October 2023, 8.6% of the bibliographic metadata stored in OpenCitations Meta are not in English. Nevertheless, the ingestion of an Anglo-Japanese dataset represents the first opportunity to test the soundness of a language-agnostic metadata crosswalk process for collecting data from multilingual sources, aiming to preserve bibliodiversity and to minimize information loss considering the constraints imposed by the OpenCitations data model, which does not allow the acceptance of multiple values in different translations for the same metadata field. The JaLC dataset is set to join OpenCitations' collections in November 2023, and it will be made available in RDF, CSV, and SCHOLIX formats. Data will be produced using open-source software and provided under a CC0 license via API services, web browsing interfaces, Figshare data dumps, and SPARQL endpoints, ensuring high interoperability, reuse, and semantic exploitation.

Moretti, A, Soricetti M. et al. 2024. Journal of Open Humanities Data,10(1): 21

PROGETTI

Ingesting JaLC Citation Data into OpenCitations: Methodology and Implementation - thesis project **Abstract**: The present thesis introduces the OpenCitations Data Sources Converter codebase expansion implemented for the ingestion of citation and bibliographic data from the Japan Link Center (JaLC), the only Japanese Registration Agency (RA) for DOI. The study starts by reviewing the citation data and how these are stored and represented in OpenCitations, and moves toward the definition of a methodology for the ingestion of JaLC data. This integration process is strictly tied to the overall OpenCitations' data ingestion workflow that led to the creation of two comprehensive open collections: OpenCitations Index for citation data, and OpenCitations Meta for bibliographic entities' metadata. Notably, JaLC is the first dataset organized according to the entity-oriented representation model to be included in the infrastructure using the newly developed workflow. JaLC represents also the initial dataset to be

integrated into OpenCitations which is predominantly non-English and uses non-Latin characters, consequently the main challenges encountered in language management and the adopted approaches to align the JaLC Data Schema with the OpenCitations Data Model (OCDM) are examined. The incorporation of JaLC data necessitated adaptations to manage JIDs, identifiers for publication venues provided by J-STAGE, thus updating the Identifier Manager module to support this new identifier type. The development of the software expansion required programming in Python and a theoretical background on the OpenCitations Data Model. The achieved software flexibility will be beneficial for OpenCitations not only for the JaLC data integration, but also for future ingestions dealing with datasets having the same representation model. Moreover, the adopted ingestion workflow for JaLC data offers a versatile framework that can be adapted and customized by researchers and institutions working with non-English data sources, thereby broadening its utility across various research contexts.

Link: https://github.com/opencitations/oc_ds_converter.git

Uncovering the Citation Landscape: Exploring OpenCitations COCI, OpenCitations Meta, and ERIH-PLUS in Social Sciences and Humanities Journals - Open Science project Questo progetto è stato realizzato come parte dell'esame finale del corso di Open Science (a.a. 2022/2023) tenuto dal professore Silvio Peroni, nel contesto del corso di laurea magistrale in Digital Humanities and Digital Knowledge dell'Università di Bologna.
L'obiettivo principale di questo progetto riguarda l'investigare la quantità e la natura di citazioni tra pubblicazioni in riviste di ambito SSH (Social Sciences and Humanities), utilizzando tre datasets di riferimento. In particolare, OpenCitations COCI per le citazioni, ERIH-PLUS per riviste di ambito SSH e OpenCitations Meta per risalire alle pubblicazioni coinvolte in citazioni. Le domande di ricerca nello specifico sono le seguenti:

- Quante citazioni (secondo COCI) coinvolgono, sia come entità citanti sia come entità citate, pubblicazioni in riviste SSH (secondo ERIH-PLUS) incluse in OpenCitations Meta?
- Quali sono le discipline che citano di più e quelle più citate?
- Quante citazioni partono e arrivano a pubblicazioni in OpenCitations Meta che non sono incluse in riviste SSH?

Tutti i materiali prodotti nell'ambito del progetto sono disponibili al seguente link: https://github.com/open-sci/2022-2023/blob/112e5ae35f8a890e4fc2ae8ff3b116085bc2369f/docs/Pika.py/material.md.

Link: https://github.com/open-sci/2022-2023/tree/112e5ae35f8a890e4fc2ae8ff3b116085bc2369f/docs/Pika.py

Network Analysis project Questo progetto è stato realizzato come parte dell'esame finale del corso di Network Analysis (a.a. 2022/2023) tenuto dal professore Saverio Giallorenzo, nel contesto del corso di laurea magistrale in Digital Humanities and Digital Knowledge dell'Università di Bologna.
Il progetto si focalizza sull'analisi comparativa tra reti sociali online e offline. Utilizzando una serie di misure di centralità e algoritmi di rilevamento delle comunità, il progetto si articola attraverso l'analisi di tre studi chiave:

- "Organization Mining Using Online Social Networks": Questo studio esplora la struttura organizzativa delle aziende attraverso l'analisi delle reti sociali dei dipendenti su Facebook, raggruppando le informazioni pubblicamente disponibili in cluster per ottenere insights sulla struttura organizzativa.
- "Learning to Discover Social Circles in Ego Networks": Analizza le reti ego-centrate (network centrati su un singolo individuo) per scoprire cerchie sociali tramite algoritmi di clustering dei nodi, utilizzando dati da Facebook che rappresentano le connessioni tra gli amici dell'ego.
- "What's in a Crowd? Analysis of Face-to-Face Behavioural Networks": Indaga le reti comportamentali face-to-face attraverso l'analisi di interazioni tra individui in eventi, utilizzando dispositivi di identificazione a radiofrequenza (RFID) per tracciare le prossimità e le interazioni.

L'obiettivo principale del progetto è dimostrare che le analisi e le misure applicate alle reti sociali online sono simili o equivalenti a quelle applicabili alle reti offline. Per fare ciò,le misure utilizzate nello studio sulle organizzazioni online sono state applicate al dataset dello studio sulle cerchie sociali negli ego-networks e viceversa. Successivamente, le misure più efficaci sono state adottate per l'analisi del dataset delle interazioni face-to-face, per valutare la loro applicabilità anche in contesti offline.
Nello studio sono stati utilizzati dataset pubblici e librerie Python come NetworkX e Matplotlib per eseguire le misure di centralità, come la centralità per grado, per vicinanza, per intermediazione, eigenvector centrality, e load centrality, oltre agli algoritmi di rilevamento delle comunità, come il clique percolation method e greedy modularity communities.

I risultati hanno evidenziato che, applicando le misure di centralità e gli algoritmi di rilevamento delle comunità, si possono identificare strutture e pattern simili nelle reti sociali sia online che offline. Ciò conferma l'ipotesi iniziale del team di ricerca sulla similitudine delle dinamiche sociali nei due contesti, nonostante le differenze nei metodi di raccolta dati e nelle piattaforme di interazione.
Tuttavia, sono state riscontrate alcune difficoltà nell'applicazione delle misure a causa della mancanza di dati o della loro inadeguatezza per alcuni tipi di analisi, come la rilevazione di comunità, evidenziando l'importanza di dataset ben strutturati e completi per analisi di rete efficaci.

Urban Scrawl - Information Modeling and Web Technologies project Nell'ambito del corso di Information Modeling and Web Technologies (a.a. 2022/2023), tenuto dal professore Fabio Vitali, come parte della prova finale, è stata sviluppata un'applicazione web per la presentazione e la lettura di articoli, arricchiti da metadati.
Urban Scrawl nello specifico è un magazine online incentrato sul mondo della street art, in cui gli articoli sono anche accessibili dalla home page attraverso una mappa realizzata con la libreria di JavaScript Leaflet. Tramite dei bottoni è possibile cambiare gli stili delle pagine HTML. In particolare, sono stati implementati uno stile di default, uno ispirato al mondo dei punkzines e uno all'Art Deco. I vari articoli sono anche stati annotati tramite dei tag in HTML per consentire la visualizzazione di vari metadati, quali persone, luoghi, opere di street art menzionate nel testo, eventi, etc.
Lo sviluppo del progetto ha richiesto principalmente l'utilizzo di HTML, CSS e JavaScript.

Link: https://martasoricetti.github.io/IMWT_project23/ | https://github.com/martasoricetti/IMWT_project23.git

Data Science project Questo progetto è stato realizzato come parte dell'esame finale del corso di Data Science (a.a. 2021/2022) tenuto dal professore Silvio Peroni, nel contesto del corso di laurea magistrale in Digital Humanities and Digital Knowledge dell'Università di Bologna.
L'obiettivo principale del progetto è lo sviluppo di un software che consenta di:

1. processare dati salvati in diversi formati (JSON e CSV) e salvarli in due database distinti, in particolare un graph e un relational database.
2. interrogare questi database simultaneamente seguendo operazioni predefinite

Le conoscenze apprese durante il corso e messe alla prova nello sviluppo del software sono state principalmente l'utilizzo di DBMSs (MySQL, BlazeGraph) per gestire rispettivamente relational e graph databases, scrittura di queries in SQL e SPARQL e in generale l'utilizzo di Python e in particolare della libreria Pandas.

Link: https://github.com/martasoricetti/data_science_project.git | https://github.com/martasoricetti/my_little_python.git

TalkWithMorandi - Interaction Media Design project Il progetto "TalkWithMorandi" è stato specificatamente elaborato per il Museo Morandi a Bologna, ospitato all'interno del MAMbo, come parte della prova finale del corso di Interaction Media Design (a.a. 2021/2022), tenuto dalle professoresse Sofia Pescarin e Simona Caraceni.
L'idea presentata nel design brief è essenzialmente la realizzazione di una guida interattiva sotto forma di un'animazione di Giorgio Morandi stesso che accompagna i visitatori attraverso le varie stanze del museo. Rispettando l'esibizione corrente, la proposta include l'installazione di totem digitali in ogni stanza e un'app mobile, accessibile tramite QR code all'ingresso del museo, entrambi progettati per arricchire la visita al museo con contenuti informativi e stimolare la riflessione personale. Il prototipo dell'esperienza è stato creato utilizzando Twine, uno strumento di storytelling.

Link: https://talkwithmorandiexperience.github.io/talkwithMorandi/ | https://github.com/TalkWithMorandiExperience/talkwithMorandi.git

The Uncrowned Kings of Sicily - Digital Text in the Humanities: Theories, Methodologies and Applications Il progetto è stato realizzato come parte dell'esame finale del corso di Digital Text in the Humanities: Theories,

Methodologies and Applications (a.a. 2021/2022) tenuto dalla professoressa Tiziana Mancinelli, nel contesto del corso di laurea magistrale in Digital Humanities and Digital Knowledge dell'Università di Bologna.
"The Uncrowned Kings of Sicily" consiste in una network analysis del romanzo di Stefania Auci, "I leoni di Sicilia". Le principali domande di ricerca sono le seguenti:

- Come cambiano le relazioni, i possedimenti e le attività della famiglia Florio attraverso le generazioni?
- I protagonisti sono effettivamente i personaggi centrali nel romanzo? A chi Stefania Auci dà più importanza nel romanzo e perchè?

Per rispondere a questi interrogativi le varie fasi della ricerca sono state:

1. annotazione del romanzo utilizzando lo schema TEI, mettendo in luce i personaggi, i luoghi e gli oggetti significativi menzionati;
2. estrazione di dati dal file XML per la creazione di due networks utilizzando Python, e in particolare la libreria beautifulsoup. Per ogni network sono stati creati due file CSV, uno per i nodi e uno per le loro connessioni;
3. visualizzazione dei networks con Gephi;
4. analisi dei networks tramite misure di centralità e algoritmi (algoritmo di Louvain in particolare) per identificare comunità, effettuate con librerie Python quali networkx e matplotlib.

Link: https://martasoricetti.github.io/the_florios/ | https://github.com/martasoricetti/the_florios.git

Songs TO Poems - Knowledge Representation and Extraction project Il progetto è stato realizzato come parte dell'esame finale del corso di Knowledge Representation and Extraction project (a.a. 2021/2022) tenuto dai professori Aldo Gangemi e Andrea Giovanni Nuzzolese, nel contesto del corso di laurea magistrale in Digital Humanities and Digital Knowledge dell'Università di Bologna.
Lo scopo principale del progetto è quello di confrontare, da un punto di vista semantico, l'album di Fabrizio De André "Non al denaro non all'amore né al cielo" e l'antologia di poesie di Edgar Lee Masters "Spoon River". Le domade di ricerca sono le seguenti:

- è possibile estrarre automaticamente i temi principali trattati in ciascuna delle nove poesie di Spoon River e nelle corrispondenti canzoni di De André?
- Quali sono le differenze?
- I risultati ottenuti dall'estrazione automatica di informazioni sono in linea con l'interpretazione che De André in persona dà del suo stesso album?

Per l'estrazione di topics è stato utilizzato BERTopic, una tecnica di topic modeling che sfrutta gli embeddings BERT e il c-TF-IDF per creare cluster densi che permettano di identificare argomenti facilmente interpretabili mantenendo al contempo le parole importanti nelle descrizioni degli argomenti. Dopo aver individuato gli argomenti principali per ogni canzone e poesia, KeyBERT è stato utilizzato per estrarre le parole chiave da ogni testo. BERTopic è stato poi impiegato nuovamente per associare ogni parola chiave a uno degli argomenti ottenuti precedentemente. Un tool online è stato sviluppato per esplorare i testi e visualizzare i risultati della ricerca.

Link: https://songstopoems.github.io/STOP/ | https://github.com/SongsTOPoems/STOP.git

Anastasija RomanLOD - Knowledge Organization and Cultural Heritage project Il progetto è stato realizzato come parte dell'esame finale del corso di Knowledge Organization and Cultural Heritage project (a.a. 2021/2022) tenuto dalla professoressa Francesca Tomasi, nel contesto del corso di laurea magistrale in Digital Humanities and Digital Knowledge dell'Università di Bologna.
"Anastasija RomanLOD" si basa sull'idea che sia possibile descrivere una persona storica a partire dalle connessioni tra la persona stessa e altre entità eterogenee, inclusi oggetti reali provenienti dai diversi ambiti delle biblioteche, degli archivi e dei musei. Per fare ciò, sono state esplorate le possibilità offerte da tecnologie del Web Semantico, come vocabolari, ontologie, RDF e URI. Pertanto, l'obiettivo del progetto è modellare un ambiente che sfrutta i linked open data per connettere concetti, oggetti, persone, luoghi, eventi e, più in generale, dati e informazioni su Anastasija Romanov, nel modo più interoperabile e semanticamente significativo possibile.
Dopo aver attentemente selezionato gli oggetti digitali da includere nella collezione, un processo graduale di mappatura, astrazione e analisi è stato condotto nel dominio di studio. Dopo una prima rappresentazione attraverso una Mappa Concettuale e un Modello E/R, sono stati analizzati e allineati i Metadata Standards utilizzati dalle diverse

istituzioni per descrivere gli oggetti. Gli ultimi due passaggi della fase di Knowledge Organization di Anastasija RomanLOD sono stati la creazione di un Theoretical Model, per espandere le relazioni già evidenziate, creare nuovi collegamenti tra i nostri oggetti e altri, e aggiungere ulteriori informazioni, e la sua formalizzazione attraverso predicati da schemi, vocabolari e ontologie già esistenti, cioè il Conceptual Model.
La fase di Knowledge Representation ha compreso la creazione di tabelle, produzione di RDF e una rappresentazione grafica dei dati RDF. Ecco un riassunto delle azioni svolte:

1. Tabelle: Vengono fornite tabelle per diversi elementi relativi ad Anastasija Romanov, ciascuna contenente triple composte da soggetto, predicato e oggetto. Mentre i soggetti e gli oggetti sono espressi in linguaggio naturale, i predicati sfruttano alcune proprietà delle ontologie.
2. Produzione di RDF: Sono forniti dati RDF per quattro elementi chiave correlati ad Anastasija Romanov: il film "Anastasia" del 1956, il ritratto della famiglia imperiale, la canzone "Once Upon a December" e il libro "Through the Russian Revolution". Ogni insieme di triple RDF fornisce informazioni strutturate su ciascun elemento, utilizzando ontologie e collegamenti semanticamente ricchi.
3. Rappresentazione RDF: Viene mostrata una rappresentazione grafica dei dati RDF relativi ai quattro elementi. La rappresentazione grafica utilizza nodi per rappresentare gli elementi e frecce per collegare soggetti, predicati e oggetti delle triple RDF.

Link: https://anastasia-romanlod.github.io/Anastasia-RomanLOD/ | https://github.com/Anastasia-RomanLOD/Anastasia-RomanLOD.git

---