UNIVERSITÀ DI PISA

Corso di Laurea in Informatica Umanistica

RELAZIONE

‘EasyScrape’

un ambiente Web per la raccolta e l'indicizzazione di articoli scientifici a supporto degli utenti delle Infrastrutture di Ricerca

Candidato: *Jacopo Gentili*

Relatore: *Angelo Mario Del Grosso*

Correlatore: *Paola Baroni*

Anno Accademico 2021-2022

INDICE

“Il peggior nemico che puoi incontrare sarai sempre tu per te stesso.”

Friedrich Wilhelm Nietzsche
Così parlò Zarathustra. Un libro per tutti e per nessuno (1883-1885)

Introduzione

Considerando il gran numero di siti da cui è possibile recuperare informazioni, la ricerca di articoli scientifici sul Web risulta spesso lunga e dispersiva. Gli studiosi solitamente impiegano molto tempo per reperire articoli e verificare la loro pertinenza con i propri interessi di ricerca. Un modo per rendere più efficienti i processi di ricerca sul Web e anche per ottimizzare il lavoro delle infrastrutture di ricerca può essere costituito da un'organizzazione sistematica dei contenuti scientifici reperibili su Internet. Questo elaborato vuole illustrare due soluzioni digitali per l'organizzazione degli articoli scientifici all'interno di un sito Web.

Le infrastrutture di ricerca rappresentano il centro nevralgico dello sviluppo e dell'innovazione sia in Europa sia nel resto del mondo, includendo ogni ambito della scienza e della ricerca. L'avvento del digitale ha lanciato sfide inedite per la ricerca, sia a livello concettuale che metodologico, stimolando ogni ambito della conoscenza a porsi domande e ad affrontare nuove sfide; ad esempio, la creazione e l'analisi di edizioni critiche digitali sono il frutto di una progressiva riflessione metodologica sugli studi testuali con il supporto di tecnologie digitali. A questa rinnovata metodologia, chiamata *Digital Scholarly Editing* è dedicato il primo capitolo dell'elaborato, dove viene fornita una panoramica degli articoli recuperati dalle soluzioni Web e viene descritto quello che può essere considerato un "caso d'uso" innovativo nella ricerca testuale nel contesto digitale, al fine di sottolineare l'apertura verso le sfide future da parte delle infrastrutture di ricerca e di sensibilizzare quindi la comunità di ricerca alle esigenze scaturite dall'interazione di più discipline. Per questo motivo nel secondo capitolo dell'elaborato si è ritenuto opportuno analizzare la struttura, il funzionamento e l'offerta di servizi e competenze di due infrastrutture di ricerca di interesse pan-europeo incentrate sulle discipline umanistiche: CLARIN ERIC, rivolta agli studiosi di Scienze Umane e Sociali, e DARIAH ERIC, rivolta agli studiosi di Arti e Scienze Umane.

Le due soluzioni digitali per l'organizzazione degli articoli scientifici all'interno di un sito Web proposte nell'elaborato sono scaturite da una riflessione sull'importanza della realizzazione, nell'ambito della ricerca scientifica, di repository di dati globali, tematici e centralizzati. La raccolta e l'organizzazione di informazioni, articoli e strumenti utili ad una determinata disciplina, infatti, possono contribuire a facilitare la ricerca di dati scientifici in contesti Web.

La prima soluzione è analizzata nel terzo capitolo e vede protagoniste due applicazioni Web ben consolidate: WordPress e Zotero. La prima consente la creazione di siti Web, prevalentemente di tipo blog. La seconda consente la creazione, online o in locale, di biblioteche digitali personali o condivise. Dopo una ricerca manuale di articoli inerenti all'ambito del *Digital Scholarly Editing*, grazie ad un'opportuna configurazione del plugin Zotpress è possibile inserirli ed indicizzarli in Zotero, per poi visualizzarli nella Zotero Library e ricercarli nella stessa tramite l'utilizzo di keyword, tag e cartelle.

La seconda soluzione - realizzata dal candidato - viene analizzata nel quarto capitolo e prevede lo sviluppo di un sito Web (denominato EasyScrape), che utilizza una tecnica di *scraping* implementata con i linguaggi PHP, HTML e XPath per recuperare i link di un gran numero di articoli scientifici sul *Digital Scholarly Editing* presenti in quattro siti Web di interesse per le discipline umanistiche. Questi link vengono estratti unitamente ai titoli, ai nomi degli autori e agli anni di pubblicazione degli articoli e vengono riorganizzati in tabelle per creare dei repository di rimandi ad articoli situati su altri siti Web, determinando così la costruzione di un "ponte di conoscenza" tra l'utente e chi offre le risorse, il quale favorisce una circolazione di informazioni più semplice e centralizzata attraverso la ricerca delle risorse all'interno del database grazie al filtro presente nella tabella o tramite il filtro avanzato.

Capitolo 1: Il *Digital Scholarly Editing*

Nel corso degli anni, l'avvento di nuove tecnologie ha profondamente cambiato i metodi e la pratica scientifica. Così anche nelle Discipline Umanistiche, le quali studiano principalmente manufatti culturali come testi, immagini o oggetti fisici conservati principalmente in biblioteche, archivi o musei. Gli studiosi spesso si trovano a lavorare con i surrogati dei manufatti originali, creati appositamente per renderli più accessibili e per facilitarne l'indagine senza comprometterne lo stato.

Non è stato facile dare una definizione di *Digital Scholarly Editing* che possa essere completa, chiara e non ambigua. Gli studiosi del settore si sono interrogati sulla natura di questo ambito e se costituisce una disciplina a sé stante oppure se rappresenta soltanto una nuova metodologia che fa uso del mezzo digitale. Per trovare una risposta è necessario analizzare i cambiamenti del paradigma di ricerca dati dalla possibilità di gestire efficacemente e di elaborare rapidamente una quantità maggiore di dati. Si rende possibile confrontare ad esempio: lo stato socio-economico degli scribi e/o commissari di manoscritti con il formato e l'impaginazione della pagina, la densità del testo e la natura o il genere dell'opera ovvero come questa si sviluppa nel tempo oppure in un'area geografica.

Grazie all'adozione del digitale, lo studioso testuale può consultare cataloghi online per reperire manoscritti e altre fonti primarie più facilmente e rapidamente di quanto non sia stato finora. Questo è stato possibile mediante l'uso consapevole di standard e protocolli per garantire la qualità e l'interoperabilità dei metadati, consentendo così di interrogare molteplici database contemporaneamente.

Una volta trovate le fonti primarie, in molti casi, è possibile visualizzarne le immagini digitali e se queste sono ad alta risoluzione possono essere addirittura più utili della consultazione dell'originale. Difatti la disponibilità di facsimili digitali rappresenta uno dei più importanti avanzamenti derivanti dall'adozione delle tecnologie digitali. La disponibilità di immagini digitali ha inoltre incoraggiato lo sviluppo della paleografia digitale e della codicologia quantitativa, nonché la ricerca sul riconoscimento automatico della grafia e sui sistemi di riconoscimento ottico dei caratteri (OCR) sia per

le edizioni sia per i manoscritti e per i primi libri a stampa. Oltre alle immagini digitali, esiste un vasto numero di versioni elettroniche di testi trascritti su tutto il patrimonio culturale, disponibili liberamente su Internet. Anche se molti di questi, purtroppo, sono quasi inutilizzabili per motivi ascrivibili ad esempio ad elaborazioni OCR non corrette, oppure alla disponibilità di vecchie edizioni senza copyright.

Da una prima riflessione, quindi, sembra che non siano cambiati solo i metodi, ma che il nuovo mezzo richieda anche un ripensamento teorico sull' impatto che esso ha sulla disciplina e sulle nozioni di testualità. In tale dibattito si inserisce anche lo sviluppo di edizioni critiche, in quanto centrale per gli studi umanistici, in quanto abbraccia quasi tutte le discipline che ne fanno parte. Nel corso del tempo, si è evoluto in un'area di ricerca indipendente che offre un ampio corpus di letteratura teorica, metodologie sofisticate, associazioni, società, conferenze, riviste dedicate e persino corsi di studio; ciò implica l'applicazione di un'ampia conoscenza, che va dalla critica materiale e bibliografica alla comprensione storica e alla critica testuale, portando spesso a forme di pubblicazione molto complesse.

1.1 La critica testuale nell'era digitale

La critica testuale è un processo che riguarda molteplici aspetti nella interpretazione, ricostituzione e pubblicazione di un testo. Si pensi ad esempio alle regole che sono applicate nella trascrizione di un documento. Mentre la trascrizione stessa è una rappresentazione, la specificazione di regole e la loro applicazione rende il processo critico, il quale identifica strutture, entità nominali e altri elementi d'interesse rendendoli espliciti; l'annotazione linguistica, ad esempio, è una forma di critica. Giudizi sulla punteggiatura, ortografia, lessico e successive sistemazioni sono i tipici ambiti della critica testuale. In più, le varie edizioni del documento devono fornire allo studioso l'autorialità e l'affidabilità dell'edizione per i suoi studi. La parola "critica" denota le attività che applicano conoscenze scientifiche giustificando il processo di riproduzione dei documenti trasformando un documento o testo in un'edizione. L'elaborazione critica del materiale è una seconda condizione necessaria per un'edizione. Infatti, una rappresentazione digitale senza alcun trattamento o senza l'aggiunta di informazioni critiche non può essere considerata un'edizione, ma tuttalpiù un facsimile, una riproduzione, un archivio digitale o una biblioteca digitale.

Le nuove metodologie testuali riguardanti il riconoscimento delle variazioni testuali trovarono, grazie all'ambiente digitale, un mezzo dove le letture varianti potevano essere presentate ai lettori in modo più efficace rispetto alla stampa, la quale ha sviluppato nel corso del tempo una modalità di presentazione delle lettere testuali profondamente insoddisfacente. Possiamo sintetizzare la critica degli apparati nei seguenti punti: è composta da formalismi abbreviati, elaborati per una tecnologia dove lo spazio è limitato, costituendo una soglia culturale solamente accessibile alle persone con un alto livello di istruzione nello specifico campo. L'indecifrabilità di molti apparati critici fa sì che i lettori spesso li ignorino tralasciandone la ricchezza analitica.

Nell'ambiente digitale, invece, l'assenza di limiti di spazio ha reso i testi molto più intuitivi, dinamici ed estendibili rispetto ai corrispondenti nella stampa, rendendo possibile la presentazione del testo sotto aspetti diversi e con funzioni diverse in base al fenomeno da presentare oltre alla possibilità di modificare l'edizione in ogni momento, anche dopo la pubblicazione. È possibile quindi individuare diversi tipi di variazione e diversi contesti e forme diverse: la prima variazione risiede nel fatto che ognuno dei lettori accederà al testo da diversi dispositivi e perciò non potrà essere controllata la forma e la dimensione del testo dagli editori. Infatti leggere un testo sullo schermo di un cellulare può essere radicalmente diverso da quello di un tablet o di un laptop; lo stesso vale per i testi letti da un browser Web o scaricati all'interno un'applicazione eReader o anche stampato su carta; la seconda forma di variazione riguarda la possibilità del mezzo digitale di apportare facili modifiche, anche dopo la pubblicazione, per questo la modifica può richiedere anche un lasso di tempo, a differenza di una situazione statica e permanente che vede la carta protagonista. Molte edizioni sono infatti pubblicate prematuramente nella loro fase di elaborazione con lo scopo di raccogliere feedback e mantenere l'interesse dei lettori per tutta la vita del progetto; la terza è una caratteristica di molte edizioni digitali, ovvero la possibilità di mostrare lo stesso testo in modi diversi, applicando insiemi di regole contenute nel cosiddetto *stylesheets*, se il testo è stato prodotto dal significato del testo codificato, è possibile visualizzare un numero diverso di formati, per esempio come un'edizione critica, diplomatica, variorum o come testo da leggere. Queste visualizzazioni possono essere generate dinamicamente, in base alla richiesta degli utenti e costituiscono il punto fermo di quello che è stato definito come edizione paradigmatica, dove la variazione di paradigmatica sta al centro dell'impostazione teorica sia dell'edizione che della sua pubblicazione.

Il mezzo digitale con le sue variazioni presenta un ambiente ideale per interfacciarsi con le instabilità testuali in modo da superare i limiti della stampa. Da quando lo spazio non è più un problema e l'ipertestualità semplifica la navigazione tra differenti testimoni e versioni di uno stesso testo, gli editori hanno accolto il nuovo formato per esplorare efficacemente le diverse letture varianti, traendone vantaggio filologico.

1.2 Edizione critica ed edizione critica digitale

Prima di addentrarsi in questo ambito è utile fornire alcune definizioni, in particolare quella di edizione critica (edizione scientificamente curata): "*L'edizione critica di un testo letterario è il risultato di una serie di operazioni condotte con metodo scientifico, ossia verificabile e dimostrato, che mirano a stabilire, secondo l'ipotesi più economica, la forma del testo più vicina possibile alla volontà dell'autore. Il concetto di testo è un concetto dinamico, dal momento che esso subisce variazioni e modifiche, sia da parte dell'autore sia da parte dei copisti durante la trasmissione nel corso del tempo."*[1]. Per dirla con le parole impiegate da Patrick Sahle nell'articolo "What is a Scholarly Digital

[1] Rossi L. C. (2003), "Finalità e metodi della filologia", ICoN Italian Culture on the Net, Modulo 276.

Edition?”, “*un'edizione scientifica è la rappresentazione critica di documenti storici*”[2]. L’autore motiva così l’utilizzo dei termini citati nella definizione: “*Rappresentazione significa ricodificare un documento o un'opera astratta e la sua trasformazione nello stesso o in un altro tipo di media. Questo di solito viene fatto sul livello visivo mediante la riproduzione dell'immagine o sul livello testuale più astratto mediante la trascrizione [...]. Critico indica un'ampia comprensione di tutti i tipi di edizioni scientifiche [...]. Inoltre la critica come pratica e come processo può assumere forme diverse. Si pensi alle regole che vengono applicate nella trascrizione di un documento. [...] La parola critica può fungere da contenitore per tutte quelle attività che applicano le conoscenze e il ragionamento critico al processo di riproduzione, di documenti e di trasformazione di un documento o di un testo in un'edizione. [...] La rappresentazione critica come nozione composita di editing, mira a ricostruire e riprodurre testi dedicandosi alla loro dimensione materiale e visiva, così come la loro dimensione astratta e intenzionale*.”[3]

Dopo aver definito cosa sia una edizione critica, è necessario definire anche cosa si intende per edizione critica digitale.

Le edizioni digitali sono diverse dalle edizioni cartacee per contenuto, struttura e funzione. Tuttavia, esse condividono lo stesso argomento e hanno gli stessi obiettivi. Per questo motivo, è possibile attenersi alla stessa definizione generale. La differenza non è tanto tra edizioni a stampa ed edizioni digitali, ma tra le varie forme di edizione. Le edizioni digitali hanno aspetti specifici, caratteristici. Alcuni di essi possono essere ottenuti trasformando le edizioni cartacee in testi elettronici e pubblicazioni digitali. Si ottengono così proprietà di accessibilità, ricercabilità, usabilità e computabilità. Con la mera digitalizzazione del materiale stampato, le caratteristiche di un cambio di paradigma centrato sul digitale non possono essere realizzate. Come sostiene Patrick Sahle, *“L'imaging digitale dei documenti originali e l’elaborazione del testo codificato in formato digitale possono essere citati come due esempi di questo fenomeno.”*[4]

Le edizioni a stampa, di solito, sono prive di facsimili come controparte visiva del testo tipografico, mentre le edizioni digitali di solito si accompagnano con le rappresentazioni visive dell’originale (edizioni *image-based*). Finché i contenuti e le funzionalità di un'edizione pensata e sviluppata con modalità tipografiche non cambiano realmente con la traduzione in formato digitale, non dovremmo chiamare queste edizioni derivate "digitali", ma “digitalizzate”. Citando ancora Patrick Sahle, *“È la struttura concettuale a definire l’edizione, non il metodo di archiviazione delle informazioni né su carta né come bit e byte.”*[5]

Tuttavia, una caratteristica principale di un'edizione digitale è la sua rappresentazione in un numero potenzialmente elevato di documenti generati dallo stesso codice secondo certe modulazioni e in un numero potenzialmente illimitato di viste diverse, talvolta

[2] Sahle, P. (2016), “2. What is a Scholarly Digital Edition?”. In Driscoll, M. J. and Pierazzo, E. (Eds.), *Digital scholarly editing: Theories and practices* (pp. 19–40), Digital Humanities Series, Vol. 4, Cambridge (UK): Open Book Publishers, p. 23.
[3] Ibid., pp. 23-24.
[4] Ibid., p. 27.
[5] Ibid., p. 27.

scelte dall'utente come ad esempio: il facsimile, la trascrizione diplomatica e le versioni di lettura. Inoltre un'edizione digitale prevede strumenti integrati per la ricerca offrendo così un più alto grado di interattività con il testo o con la gestione del layout o dei download del codice sorgente, grazie ai diversi percorsi di navigazione e alla presenza di collegamenti ipertestuali. Dunque la mera digitalizzazione non fa di un' edizione cartacea un'edizione digitale. C'è ancora la differenza nella struttura generale dell'intero paradigma. Come sostiene Patrick Sahle, tale differenza può essere descritta in modo piuttosto vago con la seguente affermazione: "*Le edizioni critiche digitali sono edizioni accademiche che sono guidate da un paradigma digitale nella loro teoria, metodo e pratica.*"[6]

Tra i vantaggi che le edizioni critiche digitali possono offrire rispetto alle edizioni cartacee spicca la possibilità di presentare e gestire quantità di dati che non sono normalmente pubblicabili in un libro stampato e instaurare connessioni attraverso l'elaborazione dei dati a velocità, precisione e completezza altrimenti non attuabile. Un'altro vantaggio è sicuramente l'interoperabilità, ovvero l'abilità di condividere informazioni in ambienti informatici grazie alla loro multimedialità e multimodalità, consentendo quindi l'organizzazione dei dati in ipertesti gerarchicamente strutturati; e infine l'interazione con l'utente.

1.3 Il paradigma digitale

I progetti per la realizzazione di edizioni critiche iniziano dalla disponibilità di facsimili digitali che successivamente vengono usati per creare le trascrizioni e le diverse versioni del testo. In contrasto con il paradigma del testo unico dell'edizione cartacea, l'edizione digitale mostra una forte tendenza verso più testi. Spesso le edizioni offrono una versione diplomatica di un testo composto trattato criticamente, altre volte i testi vengono forniti anche in traduzione estraendo le informazioni semantiche da organizzare in un database o presentate da indici. All'interno del paradigma tipografico, il testo edito è di gran lunga la caratteristica più importante, ovvero il fulcro e il centro dell'edizione. Tutte le altre informazioni di corredo, come immagini illustrative, informazioni bibliografiche, dettagli di scrittura e composizione, letture varianti o interpretazioni semantiche, sono solo substrati o fortificazioni. Invece all'interno del paradigma digitale, il processo è invertito, ovvero si sviluppa gradualmente dai documenti, dalle testimonianze visive che offre la trascrizione, attraverso l'applicazione di conoscenze critiche, storiche, stilistiche e filologiche. Inoltre le edizioni devono essere conformi agli standard per essere accettate come base per ulteriori ricerche scientifiche, oltre a fornire una rappresentazione completa di ciò che rappresentano.

Ovviamente, un'edizione critica arriva con la promessa di affidabilità e standard elevati. Immagini digitali, trascrizioni, critica testuale, commenti, annotazioni e testi contestuali devono avvalorare la rappresentazione del soggetto editoriale, curando la qualità in ambito scientifico insieme all'applicazione di tutte le conoscenze esistenti e

[6] Ibid., p. 28.

aggiungendo quindi la dimensione critica a pubblicazioni altrimenti potenzialmente acritiche con riferimento alla tradizione metodologica dell'edizione critica.

1.4 La Text Encoding Initiative

Lo sviluppo di strumenti e software in grado di elaborare i dati ha permesso di facilitare il lavoro editoriale e guidare gli studiosi verso nuovi orizzonti della ricerca. Queste nuove tecnologie richiedono la creazione di un vocabolario condiviso vista la grande varietà di testi, e della grande varietà di fenomeni editoriali, nasce così l'esigenza di utilizzare standard accettati per una corretta gestione dei metadati, oltre alla trascrizione dei testi e la descrizione di eventi, persone e date.

Questa è forse l'area in cui la ricerca ha fatto più progressi: la prima iniziativa, la *Text Encoding Initiative*, nel 1986, antecedente allo sviluppo del World Wide Web, è stata la base per lo sviluppo dell'idea stessa di *Digital Scholarly Editing*. Nonostante l'uso diffuso della TEI in tutte le fasi del processo di editing, molto resta ancora da fare. Il "problema" della TEI è che la sua completezza e flessibilità rendono difficile per gli sviluppatori creare strumenti efficaci che possano essere generalizzabili e servire più di un progetto alla volta. Tuttavia, lo sforzo verso la standardizzazione ha permesso di sviluppare una comunità internazionale e transdisciplinare interessata all'editing digitale scientificamente curato. Inoltre, l'esistenza del consorzio TEI e dello standard di codifica sta evidenziando le aree in cui la fase di standardizzazione deve ancora migliorare.

Lo scopo della *Text Encoding Initiative* è quindi quello di fornire linee guida che specificano quali metodi di codifica adottare per la creazione e la gestione in forma digitale potenzialmente di ogni tipo di dato pertinente ai ricercatori in ambito umanistico, come testi di interesse storico, manoscritti, documenti d'archivio, iscrizioni antiche e molti altri. Come suggerisce il nome, il dato testuale è quello privilegiato, ma può essere utilmente applicato a qualsiasi forma di dato digitale. La TEI insiste su ciò che è comune a ogni tipo di documento testuale, sia esso rappresentato in forma digitale su disco o scheda di memoria, in forma cartacea come libro o giornale, in forma scritta come manoscritto o codice, o incisa su pietra o su tavoletta di cera.[7] Questa continuità facilita la migrazione del testo da manifestazioni più antiche come la stampa o il manoscritto a quelle più recenti come supporti di memoria ottica o il display. Ne consegue che il modello TEI del testo è in gran parte condizionato da ciò che il testo è stato in passato, anche se prova a dare agio a tutti i tipi di documenti digitali, siano essi "nati digitali" o meno.

Attualmente i documenti TEI in forma digitale sono realizzati utilizzando un linguaggio di codifica formale a marcatori, molto diffuso, chiamato XML. Il linguaggio di markup XML è specificato, manutenuto e pubblicato dal World Wide Web Consortium (W3C) nel 1998. XML fornisce un modo semplice per rappresentare dati strutturati in un flusso

[7] Burnard, L. (2014), *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*, Marseille (FR): OpenEdition Press.

lineare di caratteri, intercalando etichette a parti specifiche del flusso. Le etichette, dette anche tag, sono denominate appositamente per indicare la funzione strutturale o la semantica delle porzioni di testo a cui si riferiscono e saranno successivamente elaborate dalla macchina. Per esempio, se i paragrafi sono chiaramente contrassegnati, un formattatore può disporli correttamente; se i toponimi sono chiaramente registrati, un programma può selezionarli automaticamente per creare un indice geografico. Tutto ciò può essere fatto in modo affidabile solo se abbiamo un certo controllo su come i tag vengono introdotti nel documento e dove questi appaiono. Un ulteriore livello di controllo è lo schema di codifica (DTD, XSD, RNG), una sorta di lessico e grammatica combinati per definire documenti XML, al fine di garantire la validità del documento stesso, rispettando non solo le regole sintattiche dello standard XML, ma anche specificando un insieme di nomi di elementi, i tipi di dati di tutti gli attributi ad essi associati e le regole sui contenuto strutturale in cui possono apparire.

Il TEI fornisce quindi schemi di codifica nei quali sono definiti centinaia di tag nonché regole su come possono essere combinati. Più esattamente, le linee guida TEI definiscono circa 600 concetti diversi, insieme a specifiche dettagliate per gli elementi XML e le classi di elementi che possono essere utilizzate per rappresentarli. La maggior parte dei documenti TEI usufruisce soltanto di un sottoinsieme di elementi tra quelli forniti; vista la modularità offerta da TEI, è opportuno scegliere solo i moduli necessari per codificare un certo soggetto editoriale e quindi soltanto i tag specifici per trattare un certo fenomeno.

Questa evoluzione degli studi testuali si dirige verso un "*upgrade*" metodologico negli studi testuali e diventa reale l'esigenza di avere un nuovo standard che sia condiviso tra tutti gli studiosi della disciplina. Per promuovere le nuove esigenze della ricerca, gli studiosi possono appoggiarsi al supporto delle infrastrutture di ricerca, le quali andranno ad integrare al loro interno i nuovi domini della ricerca e forniranno gli strumenti adatti per condurre gli studi. Saranno quindi le infrastrutture di ricerca orientate allo studio del testo e della lingua, come CLARIN ad inglobare al suo interno questo nuovo dominio della ricerca e studiarlo con i mezzi forniti dall'Unione Europea.

Capitolo 2: Le Infrastrutture di Ricerca

In questo capitolo vengono descritte le infrastrutture di ricerca di interesse pan-europeo, con un focus particolare su quelle incentrate sulle discipline umanistiche. Il contenuto di alcuni dei paragrafi successivi è tratto dal *Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027*[8], adottato mediante il Decreto Ministeriale n. 1082 del 10 settembre 2021[9].

[8] Cfr. Ministero dell'Università e della Ricerca - Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027.
[9] Cfr. Ministero dell'Università e della Ricerca - Decreto Ministeriale n. 1082 del 10 settembre 2021 - Adozione del Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027.

Le Infrastrutture di Ricerca (IR) sono strutture con sedi fisiche o virtuali che hanno la *mission* di fornire risorse e servizi alla comunità scientifica per lo svolgimento di attività di ricerca di qualità. La loro peculiarità consiste nel consentire libero accesso a tutta la comunità della ricerca, sia pubblica sia privata, sia accademica sia industriale. L'offerta delle IR include, tra le altre cose, raccolte di dati scientifici, archivi, sistemi informatici e reti di comunicazione.

La Commissione Europea (CE) definisce l'IR come *"gli impianti, le risorse e i servizi connessi utilizzati dalla comunità scientifica per compiere ricerche ad alto livello nei loro rispettivi settori e comprende i principali impianti o complessi di strumenti scientifici e il materiale di ricerca, le risorse basate sulla conoscenza quali collezioni, archivi o informazioni scientifiche strutturate e le infrastrutture basate sulle tecnologie dell'informazione e delle comunicazioni, quali le reti di tipo GRID, il materiale informatico, il software e gli strumenti di comunicazione, nonché ogni altro mezzo necessario per raggiungere il livello di eccellenza"*[10] .

Il compito della CE è garantire che le IR siano aperte e accessibili a tutti gli attori della ricerca, nel contesto comunitario come in quello extra-comunitario, fornendo loro non solo strumenti utili allo svolgimento delle attività di ricerca ma anche efficienti strategie organizzative. Gli obiettivi principali della CE riguardano lo sviluppo delle IR ed il loro sostegno, obiettivi che persegue, ad esempio, attraverso la definizione di strategie per nuove IR europee da consolidare a livello intergovernativo o nazionale, promuovendo il potenziale innovativo delle IR oppure incoraggiando la collaborazione scientifica a livello internazionale per la risoluzione di problemi comuni. La CE ha anche obiettivi secondari (ma non per questo meno importanti), tra i quali ridurre la frammentazione dell'ecosistema della ricerca ed evitare la duplicazione degli sforzi, obiettivi che persegue cercando di coordinare al meglio le varie attività delle IR.

2.1 Le Infrastrutture di Ricerca in Europa

Una fonte utile per comprendere l'importanza strategica delle IR nel contesto europeo è costituita dal PNIR 2021-2027, dove il Consorzio di Infrastrutture di Ricerca Europeo (*European Research Infrastructure Consortium* - ERIC)[11] è definito come *"una forma giuridica specifica che facilita la nascita e il funzionamento delle IR di interesse pan-europeo"*[12].

La creazione di questa forma giuridica e la conseguente formalizzazione di 22 ERIC hanno apportato un miglioramento significativo nei settori scientifici e tecnologici dello Spazio Europeo della Ricerca (*European Research Area* - ERA), non solo contribuendo alla mobilità dei ricercatori e assicurando l'accesso alle IR ma anche garantendo la

[10] Ministero dell'Università e della Ricerca - Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027, Cap. 2. Analisi dello stato dell'arte delle Infrastrutture di Ricerca.
[11] Cfr. EUR-Lex - Regolamento (CE) n. 723/2009 del Consiglio dell'Unione Europea del 25 giugno 2009, relativo al quadro giuridico comunitario applicabile ad un consorzio per un'infrastruttura europea di ricerca (ERIC).
[12] Ministero dell'Università e della Ricerca - Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027, Par. 2.1. Il contesto europeo.

diffusione e l'ottimizzazione dei risultati e facilitando l'utilizzo delle IR in programmi e progetti di ricerca.

La CE definisce, valuta e mette a disposizione strategie e strumenti per dotare l'UE di IR di livello internazionale. Contemporaneamente, coinvolge gli Stati Membri (e, in alcuni casi, certi Paesi Associati e/o Terzi) mediante vari programmi di finanziamento, rendendo così le IR aperte e accessibili. Lo scopo della combinazione delle competenze e dell'impegno dei migliori studiosi dei vari ambiti di ricerca è promuovere il potenziale d'innovazione delle IR e, allo stesso tempo, rendere l'industria più consapevole delle opportunità offerte per migliorare i loro prodotti e lo sviluppo condiviso di tecnologie avanzate. In questo contesto, si comprende quanto l'azione di coordinamento svolta dalla CE sia fondamentale al fine di evitare la duplicazione degli sforzi e ridurre la frammentazione dell'ecosistema della ricerca.

Horizon Europe (o *HORIZON*) è il programma quadro di finanziamento dell'Unione Europea (UE) per la ricerca e l'innovazione per il periodo 2021-2027. Ha una dotazione finanziaria complessiva di 95,5 miliardi di euro, di cui 5,4 destinati al piano per la ripresa *Next Generation EU*. Il processo di selezione dei progetti da finanziare è avviato tramite periodiche *call for proposals*, inviti aperti e competitivi a presentare proposte progettuali in ambito scientifico. Il programma quadro è attuato direttamente dalla CE con l'obiettivo generale di ottenere dagli investimenti comunitari una ricaduta in termini scientifici, tecnologici, economici e sociali, in modo da rafforzare le basi scientifiche e tecnologiche dell'UE, promuovere la sua competitività negli Stati Membri (rafforzando lo Spazio Europeo della Ricerca) ed attuare delle politiche europee mirate a fronteggiare le sfide globali del nostro tempo. La CE intende sfruttare al massimo il valore aggiunto offerto dall'UE, focalizzando *HORIZON* su obiettivi e attività che non possono essere realizzati in modo efficace dai singoli Stati Membri[13].

Nei paragrafi successivi vengono analizzate le varie entità giuridiche che costituiscono le IR. Prima di concentrarsi sulle IR europee incentrate sulle discipline umanistiche, è opportuno inquadrare queste realtà di ricerca nel contesto europeo, analizzando la loro struttura, le discipline di riferimento e le iniziative europee ad esse correlate.

2.1.1 ERIC: European Research Infrastructure Consortium

Nel sito ufficiale della CE l'ERIC è definito come segue:

> "*L'European Research Infrastructure Consortium (ERIC) è una forma giuridica specifica che facilita la creazione e il funzionamento di infrastrutture di ricerca di interesse europeo. Essa consente la creazione e la gestione di infrastrutture di ricerca nuove o esistenti su base non economica e può svolgere alcune attività economiche limitate legate a questo compito. L' ERIC diventa un'entità giuridica a partire*

[13] Cfr. il sito Web di APRE - Horizon Europe - Home.

dalla data di entrata in vigore della decisione della Commissione che istituisce l'ERIC."[14]

Grazie alla costituzione degli ERIC, le IR acquisiscono un riconoscimento giuridico in tutti i paesi dell'UE e la flessibilità nell'adattarsi ai requisiti specifici dei propri ambiti di competenza, incrementando in tal modo la velocità del processo della propria creazione. Le IR richiedono la partecipazione di più paesi europei, ma consentono anche quella di paesi extraeuropei. Alla comunità di ricerca di riferimento l'accesso effettivo alle IR è garantito dalle regole stabilite nell'apposito statuto. Oltre a sostenere progetti di ricerca, le IR devono contribuire sia alla diffusione e all'ottimizzazione dei risultati sia alla mobilità delle conoscenze e dei ricercatori nello Spazio Europeo della Ricerca.

2.1.2 ESFRI: European Strategy Forum on Research Infrastructures

Il Forum Strategico Europeo per le Infrastrutture di Ricerca (*European Strategy Forum on Research Infrastructures* - ESFRI) è stato istituito per indirizzare gli sforzi della ricerca verso una precisa direzione in accordo con le delegazioni nazionali degli Stati Membri. Nel periodo sotto riportato, tratto dal sito Web del Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR), viene spiegato quali sono i suoi obiettivi e come vengono stabiliti.

> *"Il Forum Strategico Europeo per le Infrastrutture di Ricerca è stato costituito nell'aprile del 2002 su mandato del Consiglio dell'Unione Europea del giugno 2001 con aggiornamenti del novembre 2004, maggio 2007 e dicembre 2012. ESFRI è composto dalle delegazioni nazionali dei 28 Stati Membri dell'Unione Europea, costituite da due rappresentanti nominati dai Ministri della ricerca, da un rappresentante della Commissione Europea, e dalle delegazioni dei Paesi Associati, attualmente dodici. [...] Il Forum contribuisce allo sviluppo di una strategia coerente per lo sviluppo delle infrastrutture di ricerca in Europa, e svolge il ruolo di incubatore agevolando le iniziative multilaterali e le negoziazioni internazionali in materia di utilizzo e sostenibilità. ESFRI realizza periodicamente la Roadmap delle infrastrutture di ricerca di dimensione pan-europea in tutti i campi della ricerca, dalle scienze fondamentali, alle scienze della vita, all'ambiente, società, integrando il patrimonio culturale ed energia. La Roadmap individua le nuove proposte di infrastruttura di ricerca, o i progetti di potenziamento di infrastrutture già attive alla luce del quadro generale degli investimenti ed è uno strumento indispensabile per facilitare il processo decisionale da parte degli Stati Membri e della Commissione Europea. In nessun modo ESFRI, che è un forum*

[14] Cfr. la pagina del sito Web della European Commission dedicata all'ERIC.

informale, solleciterà proposte specifiche o prenderà decisioni sul finanziamento e sulla localizzazione di infrastrutture future.[15]

2.1.3 Le Infrastrutture di Ricerca in Italia

In merito alle IR nel contesto italiano, nel PNIR 2021-2027 si legge:

> *"Il nostro Paese ritiene le Infrastrutture di Ricerca (IR) strategiche per lo sviluppo del sistema della ricerca nazionale. Per tale motivazione, alla strategia sulle IR è dedicato un apposito piano, il Piano Nazionale Infrastrutture di Ricerca (PNIR), parte integrante del Programma nazionale per la ricerca (PNR), previsto dal D.Lgs. 204/1998; trattasi del documento che fornisce l'orientamento strategico per le politiche della ricerca in Italia, alla realizzazione del quale contribuiscono differenti amministrazioni dello Stato, ma il cui coordinamento è in capo al Ministero dell'Università e della Ricerca.*
> *Il PNR contiene già al suo interno gli elementi principali relativi alla strategia del Paese in merito alle Infrastrutture di Ricerca, ma, anche in questa sua edizione 2021 – 2027, prevede e rimanda ad un apposito documento, il PNIR, per un maggiore dettaglio e sviluppo, e soprattutto per la definizione e aggiornamento delle priorità nazionali in tema di IR, suo compito primario."*[16]

In Italia l'organismo di finanziamento delle IR è il Ministero dell'Università e della Ricerca (MUR), che opera a stretto contatto con vari organi decisionali e di indirizzo europei e nazionali - tra cui il già citato ESFRI e l'Agenzia per la Promozione della Ricerca Europea (APRE) - ed impiega fondi europei e nazionali - come *Horizon Europe* e il Fondo Ordinario degli Enti Pubblici di Ricerca (FOE) - per sostenere le attività degli ERIC.

Il riconoscimento dell'importanza strategica delle IR nel nostro paese si è concretizzato, dal 2010 ad oggi, in uno stanziamento complessivo di oltre un miliardo di euro, attuato dal MUR tramite il FOE. Grazie a tale investimento, assegnato inizialmente su base straordinaria e, in anni recenti, strutturato stabilmente tramite la voce di finanziamento "progetti a valenza internazionale", l'Italia prende parte a 20 ERIC.

Degli ERIC partecipati dall'Italia, il presente elaborato si focalizza su quelli inerenti alle IR incentrate sulle discipline umanistiche: l'ERIC per la *Common Language Resources and Technology Infrastructure* (CLARIN ERIC) e l'ERIC per la *Digital Research Infrastructure for the Arts and Humanities* (DARIAH ERIC). CLARIN e DARIAH sono state identificate dall'ESFRI e classificate come Landmarks SSH RI nella ESFRI Roadmap del 2016.

[15] Cfr. la pagina del sito Web del Ministero dell'Istruzione, dell'Università e della Ricerca dedicata all'ESFRI.
[16] Ministero dell'Università e della Ricerca - Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027, Par. 1. Introduzione.

L'Italia partecipa a CLARIN ERIC e a DARIAH ERIC tramite il Consorzio Nazionale CLARIN-IT, coordinato dall'Istituto di Linguistica Computazionale "A. Zampolli" del Consiglio Nazionale delle Ricerche (CNR-ILC), e il Consorzio Nazionale DARIAH-IT, coordinato dal Consiglio Nazionale delle Ricerche (CNR).

2.1.4 Analisi critica

Nel PNIR 2021-2027 viene sottolineata l'importanza rivestita dalle IR per la CE, che le considera *"la spina dorsale dell'ERA riconoscendo loro il ruolo fondamentale di rendere l'Europa attraente per i migliori ricercatori di tutto il mondo, contribuendo alla condivisione della conoscenza e all'innovazione."*[17]

Viene inoltre evidenziato come nelle Conclusioni del Consiglio Europeo dell'UE del 1° dicembre 2020 sia stata ribadita *"la necessità di investire in modo sostenibile nelle IR nazionali ed europee durante tutto il loro ciclo di vita, per consentire loro di contribuire a risultati eccellenti nelle scienze fondamentali e applicate e fornire la conoscenza completa necessaria per affrontare le grandi sfide presenti e future."*[18]

Dato tale contesto, viene affermata la necessità di riflettere sul funzionamento del PNIR negli ultimi sei anni e di definire i punti di forza e le criticità del sistema infrastrutturale edificato e sostenuto tramite il FOE, *Horizon 2020* (o *H2020*)[19] ed altri fondi nazionali ed europei.

I punti di forza e le criticità individuati nel PNIR 2021-2027 sono riportati nei paragrafi seguenti.

2.1.4.1 Punti di forza

- *Qualità riconosciuta delle infrastrutture italiane di ricerca in ambito internazionale;*
- *investimenti già effettuati a sostegno della partecipazione alle principali infrastrutture di ricerca di livello europeo. Ciò consente alla comunità scientifica nazionale di avere accesso alle infrastrutture di punta localizzate in territorio europeo nei relativi campi di ricerca;*
- *partecipazione agli ERIC, con i relativi vantaggi;*
- *benefici dell'azione di potenziamento delle IR nazionali prioritarie messa in atto con il Programma Operativo Nazionale Ricerca e Innovazione 2014-2020 (PON RI 2014-2020);*
- *benefici dell'azione di rafforzamento del capitale umano delle infrastrutture di ricerca (D.D. n. 2595 del 24 dicembre 2019).*[20]

[17] Ibid., Par. 2.3 Analisi Critica.
[18] Ibid., Par. 2.3 Analisi Critica.
[19] *Horizon 2020* era il programma quadro di finanziamento dell'UE per la ricerca e l'innovazione per il periodo 2014-2020.
[20] Ministero dell'Università e della Ricerca - Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027, Par. 2.3. Analisi critica.

2.1.4.2 Criticità

- *Scarso coordinamento fra le IR, nonostante il bisogno della ricerca europea e nazionale di un'azione molto trasversale agli ambiti di ricerca;*
- *attrattività e riconoscimento del ruolo delle IR di facilitatore della ricerca, ancora da migliorare, considerando soprattutto l'impegno e le risorse necessarie alla loro creazione e operatività; è un problema non solo italiano, che porta da un lato ad un sottoutilizzo delle IR e dall'altro ad un insieme di iniziative sovrapponibili e mal coordinate fra loro;*
- *scarsa integrazione delle IR, che pur nascono nelle comunità scientifiche, in progetti di ricerca e di scambio o mobilità dei ricercatori: in H2020 le IR non sono state considerate un elemento importante per lo svolgimento di ricerche […];*
- *insufficiente coordinamento dei diversi attori (fra ministeri potenzialmente coinvolti, fra attori pubblici e privati) e delle politiche, nazionali e regionali;*
- *assenza di garanzia di una stabilità pluriennale del finanziamento, in quanto lo stesso è prevalentemente assicurato con fondi ordinari (FOE) la cui ripartizione è definita attraverso un Decreto Ministeriale con periodicità annuale;*
- *assenza di appositi strumenti di finanziamento per le IR a disposizione di altre Istituzioni pubbliche di Ricerca, differente dal FOE;*
- *assenza di un unico processo di valutazione, organico e complessivo, del panorama delle Infrastrutture di Ricerca, che comprenda anche aspetti che vadano oltre la qualità scientifica dell'IR, come ad esempio il loro impatto sulle comunità scientifiche, che vada oltre le pubblicazioni e quello socio economico, e che indirizzi il finanziamento delle IR."* [21]

Nei paragrafi successivi viene fatta una descrizione della struttura e degli obiettivi delle IR incentrate sulle discipline umanistiche, nonché della loro offerta in termini di risorse, strumenti e servizi. Entrambe danno accesso a dati linguistici (ad esempio, corpora ed edizioni critiche digitali), ma presentano altresì nuovi metodi di studio che supportano strumenti digitali, generando così nuovi domini di ricerca.

2.2 CLARIN ERIC

La *Common Language Resources and Technology Infrastructure* (CLARIN) è un'IR volta alla condivisione di una vasta gamma di dati linguistici (in forma scritta, parlata o multimodale) e strumenti avanzati per scoprire, esplorare, sfruttare, annotare, analizzare o combinare tali dati ovunque essi si trovino, dando facile accesso agli studiosi delle Scienze Umane e Sociali grazie ad una rete di Centri, distribuiti sul territorio europeo ed interconnessi, che ospitano depositi di dati linguistici digitali (o repository), strumenti avanzati ed archivi fisici o digitali all'avanguardia.

CLARIN è stata inserita nella prima pubblicazione dell'ESFRI Roadmap del 2006. Nel progetto della sua fase preparatoria (2008-2011), selezionato nell'ambito della Call

[21] Ibid., Par. 2.3 Analisi Critica.

FP7-INFRASTRUCTURES-2007-1 e finanziato con Grant Agreement n. 212230, sono state poste le basi organizzative, amministrative, tecniche e legali per la sua costituzione come IR[22]. L'istituzione di CLARIN come ERIC è stata sancita dall'approvazione del suo Statuto da parte della CE il 29 febbraio 2012[23]. CLARIN ha conseguito lo status di Landmark SSH RI nella ESFRI Roadmap del 2016.

2.2.1 Governance

L'organo di governo e di coordinamento di CLARIN è **CLARIN ERIC**, un consorzio di paesi europei rappresentati dai loro Ministeri della Ricerca.

L'organo decisionale più alto è la ***General Assembly***, composta dai rappresentanti dei Ministeri della Ricerca e/o delle Agenzie di Finanziamento dei Paesi Membri ed assistita dallo ***Scientific Advisory Board***.

La gestione quotidiana è di competenza del ***Board of Directors***, che riceve le linee guida dalla General Assembly e ad essa deve riferire. È composto dall'***Executive Director*** e da tre ***Directors*** nominati dalla General Assembly. Ed istituisce i ***Thematic Committees***[24], che lo aiutano nell'attuazione delle strategie e ad esso devono riferire.

Il coordinamento dell'attuazione delle strategie stabilite dalla General Assembly è svolto dal ***National Coordinators' Forum*** e la coerenza e la stabilità dei servizi infrastrutturali sono garantiti dallo ***Standing Committee for CLARIN Technical Centres***. Entrambi sono comitati permanenti e devono riferire al Board of Directors.

A supporto del Board of Directors, del National Coordinators' Forum, dello Standing Committee for CLARIN Technical Centres e dei Thematic Committees è chiamato il ***CLARIN Office***, che costituisce il vero motore propulsivo di CLARIN ERIC.

Il ruolo di ciascuno degli organismi che operano nell'ambito di CLARIN ERIC è dettagliato nella sezione Governance del sito Web di CLARIN ERIC[25].

2.2.2 Vision, mission, strategie e pilastri tecnici

Nella ***vision*** di CLARIN ERIC tutte le risorse e gli strumenti linguistici digitali da tutta Europa ed oltre sono accessibili attraverso un unico ambiente on-line per il supporto dei ricercatori nelle Scienze Umane e Sociali.

---

[22] Per approfondimenti, si veda la sezione About CLARIN in the Preparatory Phase del sito Web di CLARIN ERIC (https://www.clarin.eu/content/about-clarin-preparatory-phase).
[23] Lo Statuto di ERIC CLARIN descrive i diritti e gli obblighi generali, legali, governativi e finanziari di ERIC CLARIN. Lo Statuto aggiornato è stato approvato dalla CE in data 4 aprile 2018. Lo Statuto aggiornato è disponibile nella sezione CLARIN ERIC Statutes del sito Web di CLARIN ERIC (https://www.clarin.eu/content/clarin-eric-statutes).
[24] Attualmente risultano istituiti cinque Thematic Committees: il *Centre Assessment Committee*, il *Knowledge Infrastructure Committee*, il *Legal and Ethical Issues Committee*, lo *Standards Committee* e lo *User Involvement Committee*.
[25] Cfr. https://www.clarin.eu/content/governance.

La ***mission*** di CLARIN ERIC consiste quindi nel creare e mantenere un'infrastruttura per supportare la condivisione, l'uso e la sostenibilità di dati linguistici e strumenti per la ricerca nelle Scienze Umane e Sociali.

La *mission* e la *vision* sono alla base della ***strategia di lungo termine*** di CLARIN ERIC. Ogni tre anni, invece, il *Board of Directors*, in consultazione con la *General Assembly* e diversi stakeholders, sviluppa una ***strategia di medio termine***, in cui vengono presi in considerazione anche sviluppi e tendenze recenti. La *CLARIN Strategy 2021-2023* è disponibile nella sezione Vision and Strategy del sito Web di CLARIN ERIC[26].

CLARIN è un'infrastruttura digitale che offre dati, strumenti e servizi per supportare la ricerca basata sulle risorse linguistiche. I suoi ***pilastri tecnici*** sono:

- ***identità federata***: consente agli utenti di accedere a dati e servizi protetti con login e password istituzionali;
- ***identificatori persistenti***: consentono citazioni sostenibili di risorse elettroniche;
- ***archivi sostenibili***: sono archivi digitali in cui le risorse linguistiche possono essere depositate, accessibili e condivisibili;
- ***metadati flessibili e definizioni concettuali***: garantiscono l'interoperabilità semantica nella descrizione delle risorse linguistiche;
- ***ricerca di contenuti***: offre un motore di ricerca per un'ampia gamma di risorse linguistiche;
- ***concatenamento di servizi Web***: offre agli utenti la possibilità di combinare liberamente i servizi di elaborazione del linguaggio[27].

2.2.3 Consorzi Nazionali

La maggior parte delle operazioni connesse all'Infrastruttura CLARIN è attuata grazie ai *Membri* e agli *Osservatori* di CLARIN ERIC, che possono essere paesi o organizzazioni intergovernative e devono aver costituito un *Consorzio Nazionale*, tipicamente formato da organismi quali università, istituti di ricerca, biblioteche e archivi pubblici, di cui almeno uno deve avere lo status di *Centro CLARIN*. Il contributo atteso dai ***Consorzi Nazionali*** consiste nel creare e rendere accessibili raccolte di dati in lingua digitale, strumenti digitali e competenze al fine di facilitare il lavoro dei ricercatori. I *Membri* e gli *Osservatori* hanno un ampio grado di libertà nel decidere in che modo contribuire all'Infrastruttura. Nei casi in cui sia ritenuto opportuno e vantaggioso, CLARIN ERIC può stipulare accordi con *Terze Parti* (istituzioni, regioni o paesi al di fuori dell'UE) per la fornitura di competenze, servizi e risorse e/o tecnologie linguistiche.

I *Consorzi Nazionali* e le *Terze Parti* devono rispettare gli stessi criteri in merito sia alla qualità e all'interoperabilità dei dati e dei servizi offerti sia alle condizioni di accesso. L'interoperabilità è assicurata attraverso gli standard adottati nell'ambito di CLARIN.

[26] Cfr. https://www.clarin.eu/content/vision-and-strategy.
[27] Per approfondimenti, si veda la sezione CLARIN Technology: An Introduction del sito Web di CLARIN ERIC (https://www.clarin.eu/content/clarin-technology-introduction).

L'accesso ai dati e ai contenuti è in linea di principio sostenibile e conforme ai principi della Scienza Aperta.

I Coordinatori Nazionali si incontrano regolarmente nel *National Coordinators' Forum*, mentre ciascun paese Membro è rappresentato da un delegato nella *General Assembly* di CLARIN ERIC.

L'elenco aggiornato dei *Membri*, degli *Osservatori* e delle *Terze Parti* di CLARIN ERIC è disponibile nella sezione Participating Consortia del sito Web di CLARIN ERIC[28].

2.2.4 Centri

CLARIN si basa su una rete distribuita di ***Centri*** che ospitano risorse linguistiche e servizi correlati. Attualmente ce ne sono 73, per lo più aventi sede in Europa. I Centri sono raggruppati sotto i Consorzi Nazionali, ciascuno dei quali ha eletto un centro a rappresentante nello *Standing Committee for CLARIN Technical Centres* (dove si svolge la maggior parte del lavoro tecnico). La valutazione della conformità ai requisiti tecnici ed organizzativi richiesti ai candidati aspiranti allo status di Centro CLARIN è affidata da CLARIN ERIC ad un *Centre Assessment Committee* indipendente.

I *Centri* sono articolati in **6 tipologie**:

- **A-Centre (Infrastructure Centre)**: offre servizi infrastrutturali che necessitano di un alto livello di impegno (stabilità, disponibilità, persistenza); contrariamente al *B-Centre*, fornisce servizi ad altri Centri; contrariamente allo *E-Centre*, richiede l'appartenenza ad un Consorzio Nazionale;
- **B-Centre (Service Providing Centre)**: offre servizi che includono l'accesso alle risorse da esso immagazzinate e agli strumenti ad esso distribuiti tramite interfacce specifiche e conformi a CLARIN in modo stabile e persistente; contrariamente allo *A-Centre*, non fornisce servizi ad altri Centri;
- **K-Centre (Knowledge Centre)**: offre competenze e consulenze su questioni che sono rilevanti per un facile utilizzo dei servizi CLARIN da parte dei ricercatori ma non sono coperte da altri Centri;
- **E-Centre (External Centre)**: offre servizi pertinenti a CLARIN; come lo *A-Centre*, fornisce servizi ad altri Centri; contrariamente allo *A-Centre*, non richiede l'appartenenza ad un Consorzio Nazionale;
- **C-Centre (Metadata Providing Centre)**: offre metadati *'machine readable'* in modo stabile e persistente, consentendo ai fornitori di servizi di raccoglierli e di renderli navigabili, ricercabili e combinabili;
- **R-Centre (Recognized Centre)**: offre risorse e strumenti tramite siti Web standard (o servizi Web) ma, non disponendo (ancora) di fondi per partecipare all'Infrastruttura, non può assumere impegni.

Il **Centro di tipo B** è probabilmente il più ambito. È infatti integrato nell'Infrastruttura con tutti gli elementi costitutivi necessari: repository stabile, metadati strutturati,

[28] Cfr. https://www.clarin.eu/content/participating-consortia.

identificatori persistenti, accesso alle risorse protette, classificazione delle licenze, descrizioni dei servizi Web. Per ottenere la certificazione B, devono essere soddisfatti criteri molto rigorosi, *in primis* il possesso di una base tecnica ed istituzionale stabile. Attualmente ci sono 24 Centri certificati B[29] e alcune decine di candidati aspiranti alla certificazione B, come si può constatare nella panoramica dei Centri attualmente istituiti disponibile nella sezione Overview CLARIN Centres del sito Web di CLARIN ERIC[30]. Le informazioni relative a tutti i Centri sono disponibili nel CLARIN Centre Registry[31].

Il **Centro di tipo K** garantisce un trasferimento continuo di conoscenze e competenze tra gli attori coinvolti nella costruzione, nel funzionamento e nell'uso della ***CLARIN Knowledge Infrastructure***[32], il cui scopo è garantire che le conoscenze e le competenze disponibili non esistano come una raccolta disomogenea di frammenti scollegati ma siano rese accessibili in modo organizzato sia alla comunità CLARIN sia, in senso più ampio, alla comunità di ricerca delle Scienze Umane e Sociali. Una panoramica dei K-Centres attualmente istituiti è offerta nella sezione Overview of CLARIN K-Centres del sito Web di CLARIN ERIC[33].

Il **Centro di tipo C**, fornendo i metadati relativi a risorse accessibili via Web, è il minimo indispensabile. Il **Centro di tipo R**, offrendo risorse e strumenti accessibili via Web (o tramite un servizio Web) senza fornirne i metadati, può rappresentare il primo passo per diventare un C-Centre.

Le istituzioni aventi sede in paesi che fanno parte di CLARIN ERIC possono istituire Centri di tipo A, B, K e C. Le istituzioni aventi sede in paesi che non fanno parte di CLARIN ERIC possono istituire Centri di tipo K, E, C e R.

2.2.4.1 Centri di Conoscenza

Questo tipo di centro è fondamentale per la circolazione e la condivisione di conoscenze e competenze su vari temi di ricerca. Ogni K-Centre ha specifiche aree di competenza (ad esempio, famiglie linguistiche o gruppi di lingue, testi scritti con modalità diverse, argomenti linguistici, elaborazione del linguaggio, analisi del parlato, costruzione di treebank, traduzione automatica), offre alcuni tipi di dati (tra cui dati lessicali e banche terminologiche) e tratta metodi e problemi generali (come gestione dei dati, aspetti etici, IPR e OCR). Nel sito Web di ogni K-Centre imprescindibile è la sezione Helpdesk, che offre agli utenti un servizio di consulenza su argomenti specifici gestito dai ricercatori del centro in tempi brevi. I Centri di tipo K possono mettere a disposizione 'luoghi' per l'archiviazione e l'estrazione di informazioni, articoli scientifici e *best-practices,* fornire indicazioni in merito alle modalità di accesso per l'utilizzo di dati e strumenti ed offrire seminari e corsi di formazione online (su analisi basate sui dati, ad esempio).

[29] Cfr. https://www.clarin.eu/content/certified-b-centres.
[30] Cfr. https://www.clarin.eu/content/overview-clarin-centres.
[31] Cfr. https://centres.clarin.eu.
[32] Cfr. https://www.clarin.eu/content/knowledge-infrastructure.
[33] Cfr. https://vonweber.nl/cgi/kcentres_page.cgi.

2.2.5 Dati, strumenti e servizi

CLARIN mette a disposizione una vasta gamma di dati linguistici digitali che include, tra gli altri, corpora scritti e parlati, risorse multimodali e database. Queste risorse sono depositate nei vari Centri in modo distribuito. È possibile accedervi esplorando i singoli archivi oppure nei due modi offerti dall'Infrastruttura per facilitarne il reperimento:

- ***Virtual Language Observatory*** **(VLO)**[34]: aiuta ad identificare una risorsa nella sua interezza;
- ***Federated Content Search*** **(FCS)**[35]: offre un accesso per la ricerca delle risorse basato su stringhe e recupera anche le informazioni sulle citazioni delle risorse.

Per facilitare i compiti dei ricercatori linguistici, attraverso i suoi Centri CLARIN mette a disposizione anche una varietà di strumenti e servizi, tra cui la possibilità di creare archivi o collezioni virtuali di dataset. Tra questi, di particolare utilità è il ***Language Resource Switchboard***[36]*,* uno strumento che aiuta ad individuare l'applicazione Web di elaborazione del linguaggio più adeguata ai dati di interesse dell'utente.

Molti dei progetti supportati da CLARIN favoriscono una capillare condivisione di dati all'interno della comunità scientifica europea grazie ad una moltitudine di repository e di strumenti che supportano la ricerca. Il progetto *ASV Leipzig* della Leipzig University, ad esempio, oltre alla *Leipzig Corpora Collection* (LCC), che fornisce oltre 500 corpora e dizionari monolingue per più di 250 lingue, offre una raccolta di strumenti in grado di estrarre ed esplorare il testo utilizzando algoritmi per l'analisi statistica del linguaggio, per l'analisi morfologica e per l'estrazione terminologica.

2.2.5.1 Famiglie di risorse

L'iniziativa ***CLARIN Resource Families***[37] offre una panoramica *'user-friendly'* per tipo di dati delle risorse linguistiche disponibili nell'Infrastruttura CLARIN per i ricercatori delle Scienze Umane Digitali, delle Scienze Sociali e delle Tecnologie del Linguaggio Umano. Le panoramiche hanno lo scopo di facilitare la ricerca comparativa e gli elenchi sono ordinati per lingua. Gli elenchi di ciascuna famiglia di risorse includono i metadati più importanti, brevi descrizioni (dimensioni delle risorse, fonti di testo, periodi di tempo, annotazioni e licenze) e collegamenti a pagine di download e ad applicativi per la costruzione automatica delle concordanze. Gli elenchi forniscono anche collegamenti ipertestuali ad altri materiali rilevanti (come i seminari e i tutorial tematici di CLARIN e le relative videoconferenze) e un elenco di pubblicazioni chiave sulle risorse esaminate.

Attualmente le famiglie di risorse linguistiche di CLARIN includono:

- 14 famiglie di corpora: corpora di comunicazione mediata dal computer; corpora di testi accademici; corpora storici; corpora degli studenti L2; corpora legali;

[34] Cfr. https://www.clarin.eu/content/virtual-language-observatory-vlo e https://www.clarin-it.it/en/content/virtual-language-observatory.
[35] Cfr. https://www.clarin.eu/content/federated-content-search-clarin-fcs-technical-details e https://www.clarin-it.it/en/content/federated-content-search.
[36] Cfr. https://www.clarin.eu/content/language-resource-switchboard.
[37] Cfr. https://www.clarin.eu/resource-families.

corpora letterari; corpora annotati manualmente; corpora multimodali; corpora di giornali; corpora paralleli; corpora parlamentari; corpora di riferimenti; risorse per la lingua dei segni; corpora parlati;
- 6 famiglie di risorse lessicali: modelli linguistici; lessico; dizionari; risorse concettuali; glossari; elenchi di parole;
- 4 famiglie di strumenti: strumenti per la normalizzazione; strumenti per il riconoscimento di entità nominali; strumenti per la codifica e la lemmatizzazione di parti del discorso; strumenti per l'analisi del sentimento.

Accanto alla panoramica delle risorse linguistiche disponibili in CLARIN, viene fornita una panoramica di altre preziose risorse linguistiche esistenti che non sono state ancora integrate nell'Infrastruttura.

2.2.6 Progetti

Nel 2015 CLARIN ERIC ha iniziato a partecipare - in qualità di partner o coordinatore oppure attraverso i suoi nodi nazionali - a progetti strategici rilevanti per la sua *vision* e la sua *mission*. Di questi progetti, finanziati nell'ambito di *Horizon 2020*, di *Horizon Europe* e di altri programmi di finanziamento comunitari e nazionali, 11 sono terminati e 7 sono in corso: ERIC Forum, TRIPLE, ENRIITC, UPSKILLS, EOSC Future, EOSC Focus e FAIRCORE4EOSC[38].

Tra i progetti in corso, vale la pena menzionare:

- **FAIRCORE4EOSC** (*Core Components Supporting a FAIR EOSC*), focalizzato sullo sviluppo e sulla realizzazione di componenti EOSC-Core per abilitare un ecosistema FAIR EOSC[39];
- **TRIPLE** (*Transforming Research through Innovative Practices for Linked interdisciplinary Exploration*), concepito per la creare un ambiente integrato di strumenti innovativi e servizi avanzati per le Scienze Umane e Sociali[40].

Tra i progetti terminati vale la pena menzionare:

- **SSHOC** (*Social Sciences & Humanities Open Cloud*), volto ad integrare le varie infrastrutture del settore delle Scienze Umane e Sociali in un unico ambiente virtuale per favorire ricerche di alta qualità e realizzare la visione dell'Open Science Cloud nel settore delle Scienze Umane e Sociali[41];

[38] Per approfondimenti, si vedano le sezioni CLARIN in EU Projects (https://www.clarin.eu/content/clarin-eu-projects) e CLARIN in past EU Projects (https://www.clarin.eu/content/clarin-past-eu-projects) del sito Web di CLARIN ERIC.
[39] FAIRCORE4EOSC è stato selezionato nell'ambito della Call HORIZON-INFRA-2021-EOSC-01 ed è finanziato con Grant Agreement n. 101057264. A questo progetto CLARIN ERIC partecipa come partner.
[40] TRIPLE è stato selezionato nell'ambito della Call H2020-INFRAEOSC-2019-1 ed è finanziato con Grant Agreement n. 863420. A questo progetto CLARIN ERIC partecipa come partner accanto al CNR-ILC, Coordinatore Nazionale di CLARIN-IT.
[41] SSHOC è stato selezionato nell'ambito della Call H2020-INFRAEOSC-2018-2 ed è finanziato con Grant Agreement n. 823782. A questo progetto CLARIN ERIC ha partecipato come partner accanto al CNR-ILC e ad altri sei Istituti del CNR.

- **ELEXIS** (*European Lexicographic Infrastructure*), mirato a coniugare i settori delle Tecnologie Linguistiche, delle Digital Humanities e della Lessicografia Computazionale[42].

Tali progetti vedono una massiccia partecipazione di nodi nazionali, sia di CLARIN sia di altre IR del settore, garantendone la sostenibilità negli anni futuri.

2.2.6.1 Supporto a progetti ed eventi

Attraverso i suoi nodi nazionali, CLARIN ERIC supporta progetti ed eventi del settore delle Scienze Umane e Sociali. Tra essi vale la pena menzionare:

- **Kretikai Politeiai: Le istituzioni cretesi dal VII al I secolo a.C.**: progetto di ricerca dottorale che ha prodotto il dataset *Cretan Institutional Inscriptions* e una Web application per interagire con esso[43];
- **AIUCD 2021 - *DH per la società: e-qualità, partecipazione, diritti e valori nell'Era Digitale*** (Edizione Virtuale, 19-22/01/2021): 10° Congresso Annuale dell'Associazione Italiana per l'Informatica Umanistica e la Cultura Digitale (AIUCD), co-organizzato dal CNR-ILC e dal Consorzio Nazionale CLARIN-IT con il supporto di CLARIN ERIC[44];
- **EUPORIA 2021**: ciclo di webinar in materia di annotazione di testi letterari e documentari tramite linguaggi specifici di dominio, organizzato come contributo a CLARIN dal Laboratorio di Filologia Collaborativa e Cooperativa (CoPhiLab) del CNR-ILC con il supporto di CLARIN-IT[45].

Un articolo sul dataset *Cretan Institutional Inscriptions* è stato presentato alla *CLARIN Annual Conference 2021* (Virtual Edition, 27-29/09/2021) e pubblicato nei *Proceeding* ad essa correlati[46]. Nel 2022 una versione estesa dell'articolo è stata selezionata per la pubblicazione nel volume *Selected Papers from the CLARIN Annual Conference 2021*[47].

---

[42] ELEXIS è stato selezionato nell'ambito della Call H2020-INFRAIA-2016-2017 ed è finanziato con Grant Agreement n. 731015. A questo progetto CLARIN ERIC ha partecipato attraverso il CNR-ILC.
[43] Il progetto è stato condotto dalla Dott.ssa Irene Vagionakis presso il Dipartimento di Studi Umanistici dell'Università Ca' Foscari di Venezia nel periodo 2016-2019 (cfr. http://dspace.unive.it/handle/10579/17819). Il sito Web dedicato al dataset è ospitato dal Consorzio Nazionale CLARIN-IT (cfr. https://www.clarin-it.it/cretaninscriptions), mentre la Web application per interagire con il dataset è ospitata dal CLARIN B-Centre ILC4CLARIN (cfr. https://ilc4clarin.ilc.cnr.it/cretaninscriptions). Le schede descrittive del dataset e della Web application sono disponibili sotto licenza gratuita nel Repository del CLARIN B-Centre ILC4CLARIN (cfr. http://hdl.handle.net/20.500.11752/OPEN-548 e http://hdl.handle.net/20.500.11752/OPEN-550).
[44] Originariamente programmato a Pisa, a causa delle dinamiche del Coronavirus l'evento è stato tenuto in forma virtuale (cfr. https://aiucd2021.labcd.unipi.it).
[45] Il ciclo di webinar è stato tenuto tra novembre 2020 e marzo 2021 (cfr. https://cophilab.ilc.cnr.it/euporia2021).
[46] Cfr. Vagionakis I., Del Gratta R., Boschetti F., Baroni P., Del Grosso A. M., Mancinelli T. and Monachini M. (2021), " 'Cretan Institutional Inscriptions' Meets CLARIN-IT". In Monachini, M. and Eskevich M. (Eds.), *CLARIN Annual Conference Proceedings 2021* (pp. 48-53), Utrecht (NL): CLARIN ERIC.
[47] Cfr. Vagionakis, I., Del Gratta R., Boschetti F., Baroni P., Del Grosso A. M., Mancinelli T. and Monachini, M. (2022), " 'Cretan Institutional Inscriptions' Meets CLARIN-IT". In Monachini, M. and Eskevich M. (Eds.), *Selected Papers from the CLARIN Annual Conference 2021* (pp. 139-150), Linköping Electronic Conference Proceedings Series, Vol. 189, Linköping (SE): Linköping University Electronic Press.

2.2.7 Strumenti di finanziamento e di supporto

CLARIN ERIC mette a disposizione di ricercatori e docenti della comunità CLARIN strumenti di finanziamento e di sostegno mirati ad affrontare le priorità strategiche che richiedono la collaborazione transnazionale, lo scambio di competenze, la formazione o la mobilità.

- ***CLARIN Resource Families Project Funding***: strumento di finanziamento per piccoli progetti che potrebbero aiutare a migliorare il processo di cura di risorse e strumenti linguistici.
- ***CLARIN Workshop Funding***: strumento di finanziamento per workshop su un argomento in linea con una o più priorità strategiche di CLARIN.
- ***CLARIN User Involvement Funding***: strumento di finanziamento per master class, tutorial, seminari, scuole estive e materiali di formazione on-line incentrati sull'insegnamento dell'utilizzo delle risorse, degli strumenti o dei servizi di deposito di CLARIN.
- ***Teaching with CLARIN***: speciale bando volto a riconoscere e mostrare gli sforzi compiuti da insegnanti, docenti e formatori nell'ambito di tutta la rete CLARIN per soddisfare le attuali esigenze formative.
- ***CLARIN Trainer Network Programme***: bando per partecipare ad eventi di formazione presso importanti scuole estive, conferenze, azioni COST ecc. in discipline e comunità rilevanti per CLARIN.
- ***CLARIN Training Suite***: bando per contribuire alla creazione di raccolte di materiali di formazione che possono essere riutilizzati ed adattati dai docenti nella rete CLARIN in contesti educativi sia formali che informali.
- ***CLARIN Mobility Grants***: sovvenzioni per imparare ad utilizzare una risorsa o uno strumento sviluppato in un centro CLARIN o ad integrare una risorsa o uno strumento sviluppato nell'infrastruttura CLARIN o per avviare uno scambio di insegnanti e un programma di formazione per insegnanti.

Alla comunità di riferimento CLARIN-ERIC offre anche supporto pratico e finanziario per la preparazione di proposte progettuali comunitarie e diversi tipi di supporto per l'organizzazione di eventi virtuali.

- ***Support for EU-Funded Projects***: supporto pratico ai Membri di CLARIN nella preparazione di proposte progettuali su argomenti di rilevanza strategica per CLARIN da presentare nell'ambito di Horizon Europe e di altri Programmi di Finanziamento comunitari.
- ***CLARIN Virtual Events Support***: supporto pratico agli organizzatori di eventi virtuali che sono rappresentanti dei Membri o degli Osservatori di CLARIN o affiliati ai Centri CLARIN.
- ***CLARIN Seed Grants***: supporto finanziario per la preparazione di domande di finanziamento di progetti nell'ambito di Horizon Europe (all'interno del Pilastro I, ERC incluso, e del Pilastro II).

Ulteriori dettagli sugli strumenti di finanziamento e di supporto attualmente attivi sono disponibili nella sezione CLARIN Funding Hub del sito Web di CLARIN ERIC[48].

L'Infrastruttura CLARIN fornisce a docenti e ricercatori risorse inerenti alle discipline del settore delle Scienze Umane e Sociali. Gli insegnanti possono condividere materiali per la formazione e riutilizzare quelli di altri colleghi. Tra le risorse disponibili ci sono i corsi e il materiale didattico raccolti tramite il bando *Teaching with CLARIN*. Dal 2016 è inoltre disponibile una biblioteca digitale dedicata alle conferenze annuali di CLARIN e ai tutorial degli eventi formativi ed accademici di CLARIN che si configura come un utile repository di video-lezioni ad accesso libero e gratuito.

In aggiunta, è disponibile una piattaforma per la condivisione di un registro contenente le informazioni relative ai corsi di Digital Humanities sostenuti da varie organizzazioni accademiche europee. All'interno di questo database è possibile effettuare ricerche sui domini delle discipline umanistiche, sulle informazioni topografiche correlate, per un certo intervallo temporale, secondo il titolo accademico d'interesse oppure utilizzando TaDiRAH[49], una tassonomia delle attività di ricerca digitale nel settore delle Scienze Umane. La piattaforma è il frutto dell'unione di due IR europee: CLARIN e DARIAH.

2.2.8 Conferenza Annuale, 'Tour de CLARIN' e 'Best-Practice Papers'

Per coloro che lavorano alla costruzione e al funzionamento di CLARIN in tutta Europa e per i rappresentanti delle comunità d'uso nelle Scienze Umane e Sociali la ***CLARIN Annual Conference***[50] rappresenta l'evento più importante. Viene organizzata allo scopo di scambiare esperienze e *'best-practices'* di lavoro con l'Infrastruttura CLARIN e di condividere piani per sviluppi futuri. Il suo programma comprende una vasta gamma di argomenti, tra cui: la progettazione, la costruzione e il funzionamento dell'Infrastruttura CLARIN; i dati, gli strumenti e i servizi che essa contiene o dovrebbe contenere; il suo effettivo utilizzo da parte di ricercatori, insegnanti o parti interessate; la sua relazione con altri infrastrutture e progetti e con la *CLARIN Knowledge Infrastructure*. A partire dal 2012, ne sono state organizzate 11 edizioni, di cui 2 in modalità virtuale (nel 2020 e nel 2021) e 1 in modalità ibrida (nel 2022).

Per dare visibilità al vasto e variegato ecosistema di CLARIN, nel 2017 è stato istituito il ***Tour de CLARIN***[51], un'iniziativa volta a mettere periodicamente in evidenza le attività di coinvolgimento degli utenti svolte nei *Consorzi Nazionali*. I momenti salienti presentati includono: risorse e strumenti sviluppati nell'ambito dei Consorzi Nazionali, casi d'uso di cui questi ultimi sono particolarmente orgogliosi, eventi organizzati per avvicinare CLARIN alla comunità delle Scienze Umane e Sociali ed interviste a studiosi che, per le loro ricerche, utilizzano i nodi nazionali dell'Infrastruttura. Una parte del *Tour* è stata successivamente dedicata al lavoro svolto nei *Centri di tipo K* (dal 2019) e nei *Centri di tipo B* (dal 2021). I momenti salienti presentati includono: la

[48] Cfr. https://www.clarin.eu/funding.
[49] Cfr. la home page del sito Web di TaDiRaH - The Taxonomy of Digital Research Activities in the Humanities.
[50] Cfr. https://www.clarin.eu/content/clarin-annual-conference.
[51] Cfr. https://www.clarin.eu/Tour-de-CLARIN.

descrizione di ciò che i Centri offrono alla comunità di riferimento ed interviste a studiosi che collaborano con i Centri o a sviluppatori che fanno parte dei Centri. I contributi presentati nelle varie tappe del *Tour* vengono raccolti nei volumi della serie *Tour de CLARIN*, pubblicati con cadenza annuale.

Una recente iniziativa rivolta ai ricercatori della comunità CLARIN è costituita dalla ***CLARIN Collection of Best-Practice Papers***. Tale raccolta comprende articoli relativi alle *'best-practices'* su vari argomenti ed è accessibile in una cartella dedicata della CLARIN Zotero Library[52]. Gli articoli più rilevanti per la comunità CLARIN vengono selezionati tra quelli a cui si fa riferimento nei siti Web e nelle Zotero Libraries dei K-Centres. Per comodità, agli articoli vengono aggiunti dei tag (modalità, K-Centre di riferimento, parole chiave). La raccolta è un'iniziativa del *Knowledge Infrastructure Committee* ed è attuata in collaborazione con i K-Centres (che forniscono gli articoli) e con il *CLARIN Office* (che cura la raccolta).

2.3 DARIAH ERIC

L'infrastruttura di ricerca digitale DARIAH (*Digital Research Infrastructure for the Arts and Humanities*) nasce per sostenere e formare gli studiosi nelle arti e nelle discipline umanistiche, offrendo loro competenze, conoscenze, metodi, strumenti, tecnologie e servizi. L'IR sviluppa e gestisce pratiche di ricerca incentrate su tecniche informatiche, sostenendo i ricercatori nel loro utilizzo finalizzato a costruire, analizzare e interpretare le risorse digitali. DARIAH riunisce singole attività artistiche o inerenti alle Digital Humanities, condivide i risultati delle ricerche e garantisce il rispetto delle migliori pratiche sugli standard metodologici e tecnici. Attraverso la condivisione di dati, servizi e strumenti e l'offerta di varie opportunità formative, l'IR consente un'organizzazione mirata alle emergenti esigenze della ricerca, favorendo approcci transnazionali e transdisciplinari. Mediante tali attività, DARIAH promuove lo sviluppo ulteriore dei metodi di ricerca, documentando lo stato dell'arte e sostenendo la conservazione dei dati di ricerca con particolare attenzione ai diversi aspetti - come la diversità, la provenienza e la granularità - delle collezioni multimediali.

DARIAH è stata inserita nella prima pubblicazione dell'ESFRI Roadmap del 2006. Il progetto per la sua fase preparatoria (2008-2011) è stato selezionato nell'ambito della Call FP7-INFRASTRUCTURES-2007-1 e finanziato con Grant Agreement n. 211583. A febbraio 2011 DARIAH è entrata in una fase di transizione in cui sono state poste le basi per la sua costituzione come IR. L'approvazione dello Statuto da parte della CE ad agosto 2014 ha sancito l'istituzione di DARIAH come ERIC, che inizialmente contava 15 *Membri Fondatori* (Austria, Belgio, Croazia, Cipro, Danimarca, Francia, Germania, Grecia, Irlanda, Italia, Lussemburgo, Malta, Paesi Bassi, Slovenia e Serbia). DARIAH ha conseguito lo status di Landmark SSH RI nella ESFRI Roadmap del 2016. L'IR ha completato la fase di implementazione ad agosto 2019 e ha avviato la fase operativa a

[52] Cfr. https://www.zotero.org/groups/562080/clarin/library.

settembre 2019. Oggi conta 20 *Membri*, 1 *Osservatore* e diversi *Partner Cooperanti* in 6 paesi non membri ed è considerata un hub paneuropeo di eccellenza scientifica[53].

2.3.1 Vision, mission e strategia

La ***vision*** di DARIAH pone le arti e le discipline umanistiche al centro di una società della conoscenza tecnologicamente in evoluzione.

La ***mission*** di DARIAH è quella di incrementare il potenziale della comunità di ricerca fornendo metodi all'avanguardia che consentano di creare, connettere e condividere conoscenze in merito alla cultura e alla società.

Le principali aree di attività dell'IR sono mirate a garantire che i ricercatori umanistici possano: i) valutare l'impatto della tecnologia sul proprio lavoro in modo informato; ii) accedere ai dati, agli strumenti, ai servizi, alle conoscenze e alle reti di cui hanno bisogno senza soluzione di continuità e in ambienti virtuali e umani ricchi di contesto; iii) attivare borse di studio abilitate digitalmente, riutilizzabili, visibili e sostenibili.

Le sfide che emergono dall'intersezione dei settori e dei metodi di ricerca consolidati con la tecnologia e il progresso tecnologico sono al centro della ***strategia*** di DARIAH, che poggia su **4 pilastri**:

- **1° pilastro**: riguarda la realizzazione di un 'mercato' per facilitare lo scambio di informazioni garantendo l'accesso a risorse ottimizzate per i ricercatori; questo punto di accesso è stato sviluppato come componente del Cloud Europeo per la Scienza Aperta (*European Open Science Cloud* - EOSC);
- **2° pilastro**: è mirato a dare accesso ad istruzione e formazione grazie a servizi come *DARIAH-CAMPUS*, *DARIAH Teach*, *PARTHENOS Training* e *DH Course Registry*[54];
- **3° pilastro**: è relativo alla costituzione di Working Groups, Hubs ed altre forme di organizzazione Transnazionale e Transdisciplinare;
- **4° pilastro**: è incentrato sulla costruzione di 'ponti' tra i *policy makers* europei e la comunità di ricerca che crea o utilizza strumenti, metodi e servizi digitali nelle arti e nelle discipline umanistiche.

I suddetti pilastri derivano dalla *mission* e dalla *vision* di DARIAH e rappresentano le priorità organizzative dell'IR ed i servizi da essa forniti alla comunità di riferimento[55].

[53] La mappa aggiornata dei *Membri*, degli *Osservatori* e dei *Partner Cooperanti* di DARIAH ERIC è disponibile nella sezione Members and Partners del sito Web di DARIAH-EU (https://www.dariah.eu/network/members-and-partners).
[54] Cfr. la sezione 2.3.5.
[55] Per approfondimenti, si veda la sezione Mission & Vision del sito Web di DARIAH-EU (https://www.dariah.eu/about/mission-vision).

2.3.2 Centri di Competenza Virtuali

L'IR opera attraverso le reti europee dei suoi ***Virtual Competence Centers*** (VCC) e dei *Working Groups* ad essi associati[56]. I VCC sono interdisciplinari, multi-istituzionali ed internazionali, ciascuno con una specifica area di competenza:

- **VCC1**: gestisce un'ambiente digitale per la condivisione di dati e strumenti sviluppati dalla comunità e garantisce la qualità, la permanenza e la crescita dei servizi tecnici per le arti e le discipline umanistiche;
- **VCC2**: mette in collegamento la ricerca e l'istruzione;
- **VCC3**: gestisce i contenuti scientifici nelle varie fasi: dalla creazione alla cura delle risorse digitali fino alla loro diffusione e condivisione e ai risultati dedicati al riutilizzo;
- **VCC4**: si concentra sulla ricerca di sponsor che sostengano le ricerche da un punto di vista finanziario e promozionale.

Sostenendo lo sviluppo sostenibile della ricerca relativa al digitale nell'ambito delle arti e delle discipline umanistiche e realizzando servizi che aiutano i ricercatori che lavorano con metodi digitali a far avanzare ulteriormente le loro ricerche, DARIAH ha impatto su 4 domini interconnessi: ricerca, istruzione, cultura ed economia. Essa, inoltre, offre materiale didattico ed opportunità di insegnamento per sviluppare competenze di ricerca digitale.

2.3.3 Scienza Aperta

La ricerca di libero accesso è sempre stata importante per DARIAH, come dimostrano i collaborativi e innovativi mezzi di ricerca digitale da essa offerti. I membri dell'IR sostengono che l'*Open Science* sia più performante se collocata tra diverse discipline. Uno degli obiettivi, pertanto, consiste nel colmare il divario tra i principi e i valori fondamentali della Scienza Aperta (quali la trasparenza, l'uguaglianza, l'innovazione e la giustizia socio-cognitiva nella creazione di conoscenza) e le pratiche comunitarie in una vasta varietà di ambiti. 'Open Science' si riferisce quindi alla libera diffusione dei risultati delle ricerche scientifiche e lo scopo della disciplina è rendere libero ogni step della ricerca scientifica, consentendo di accedervi ai cittadini con qualunque grado di istruzione: dagli esperti ai normali cittadini.

2.3.4 Progetti

DARIAH ERIC partecipa o ha partecipato - in qualità di coordinatore, partner o partner affiliato oppure attraverso i suoi nodi nazionali - a vari progetti strategici rilevanti per la sua *vision* e la sua *mission*. Di questi progetti, finanziati nell'ambito di *Horizon 2020*, di *ERASMUS+* e di altri programmi di finanziamento comunitari e nazionali, 15 sono

[56] Per approfondimenti, si veda la sezione Organisation and Governance del sito Web di DARIAH-EU (https://www.dariah.eu/about/organisation-and-governance).

terminati e 5 sono in corso: ERIC Forum, TRIPLE, EOSC Future, CLS INFRA e OPERAS-PLUS[57].

Tra i progetti in corso, vale la pena menzionare:

- **EOSC Future**, mirato ad integrare, consolidare e connettere *e-infrastructures*, comunità di ricerca ed iniziative per l'Open Science per implementare il Cloud Europeo per la Scienza Aperta mediante un ulteriore sviluppo del portale EOSC e dei componenti EOSC-Core e EOSC-Exchange[58];
- **CLS INFRA** (*Computational Literary Studies Infrastructure*), mirato a costruire un'infrastruttura condivisa che offra dati, strumenti e conoscenze di alta qualità per intraprendere studi letterari nell'era digitale; la stabilità e la sostenibilità a lungo termine del progetto sono garantite all'integrazione di CLARIN ERIC e DARIAH ERIC[59].

Tra i progetti terminati, **DiXiT** (*Digital Scholarly Editions Initial Training Network*) ha avuto una gran risonanza nel mondo del *Digital Scholarly Editing*. Esso ha visto la partecipazione di una rete internazionale di istituzioni coinvolte nella creazione e nella pubblicazione di edizioni critiche digitali. Era incentrato sull'offerta di un programma coordinato di formazione e ricerca sia per ricercatori in fase iniziale sia per ricercatori esperti nella multidisciplinarietà, nelle tecnologie, nelle teorie e nei metodi del *Digital Scholarly Editing*. A tal fine, ha ospitato e sostenuto una moltitudine di conferenze e workshop in tutta Europa[60].

2.3.5 Servizi

DARIAH offre alcuni servizi editoriali costituiti da repository di pubblicazioni:

- ***Hypotheses***: ospita blog di ricerca, copre tutte le aree del settore delle Scienze Umane e Sociali e offre testi in un'ampia gamma di lingue e ad accesso libero;
- ***Bibliography "Doing Digital Humanities"***: raccoglie voci bibliografiche inerenti alle Digital Humanities;
- **"*DARIAH-DE Working Papers***: presenta documenti che includono contributi di vario genere forniti nell'ambito della ricerca; tali contributi sono soggetti ad un ben definito processo di garanzia della qualità, che include il supporto editoriale attraverso un apposito team; tutti i contributi sono pubblicati in *Open Access* con licenza CC-BY.

[57] Per approfondimenti, si veda la sezione Projects del sito Web di DARIAH-EU (https://www.dariah.eu/activities/projects-list).
[58] EOSC Future è stato selezionato nell'ambito della Call Call INFRAEOSC-03-2020 ed è finanziato con Grant Agreement n. 101017536.
[59] CLS INFRA è stato selezionato nell'ambito della Call H2020-INFRAIA-2018-2020 ed è finanziato con Grant Agreement n. 101004984.
[60] DiXiT è stato selezionato nell'ambito della Call FP7-PEOPLE ed è stato finanziato con Grant Agreement n. 317436.

2.3.6 Formazione

La formazione e l'istruzione rientrano tra gli obiettivi principali dell'Infrastruttura, che predilige l'uso di nuovi metodi e ritiene il digitale lo strumento per valorizzare la ricerca umanistica di alto livello.

Uno dei servizi principali offerti dall'Infrastruttura è la piattaforma ***DARIAH-Campus***, che consente di trovare offerte di formazione e istruzione affiliate a DARIAH e riunisce al suo interno tre piattaforme:

- ***DARIAH Teach***: mette a disposizione materiali didattici (in alcuni casi, gratuiti) realizzati dalla comunità in lingua diverse;
- ***PARTHENOS Training***: offre lezioni ed esercitazioni su un'ampia varietà di argomenti;
- ***DH Course Registry***: riunisce i corsi disponibili in ambito umanistico ad ogni livello ed indica le sedi degli istituti europei in cui sono stati tenuti determinati corsi.

DARIAH promuove lo sviluppo e la consapevolezza delle competenze al di fuori delle qualifiche formali, integrando così l'istruzione fornita dal partenariato universitario.

2.4 Considerazioni finali

Come è stato detto in questo capitolo, le IR sono un elemento cruciale per lo sviluppo e l'innovazione. La loro struttura è generalmente organizzata in entità giuridiche formate da centri di ricerca, enti nazionali ed europei, università ed altri organismi incentrati sulla ricerca. Le IR si basano sullo scambio libero di informazioni (risultati di ricerche, dati, articoli scientifici, strumenti ecc.) al fine di condividere la conoscenza in maniera estensiva. Esse, inoltre, prevedono un'offerta formativa volta a supportare i ricercatori nell'utilizzo degli standard e delle metodologie migliori di un determinato settore. Nei capitoli seguenti verranno descritte due soluzioni Web per raccolta di articoli scientifici inerenti al *Digital Scholarly Editing* in un unico 'serbatoio'.

Capitolo 3: La Biblioteca Digitale del CLARIN K-Centre 'DiPText-KC'

In questo capitolo viene descritto il processo che ha portato alla realizzazione della prima release della Zotero Library del DiPText-KC (*CLARIN Knowledge Centre for Digital and Public Textual Scholarship*)[61], uno dei centri CLARIN nati sotto l'egida di CLARIN-IT[62], il nodo italiano di CLARIN. DiPText-KC è un CLARIN K-Centre che

[61] Cfr. https://diptext-kc.clarin-it.it.

[62] Gli altri centri CLARIN nati sotto l'egida di CLARIN-IT sono: i) CLARIN B-Centre ILC4CLARIN (*ILC4CLARIN Centre at the Institute for Computational Linguistics*) (cfr. http://ilc4clarin.ilc.cnr.it); ii) CLARIN C-Centre ERCC (*Eurac Research CLARIN Centre*) (cfr. http://clarin.eurac.edu); iii) CLARIN K-Centre CKCMC (*Knowledge Centre for Computer-Mediated Communication and Social Media Corpora*) (cfr. https://cmc-corpora.org/ckcmc).

opera come centro virtuale distribuito ed è costituito da due partner: il Venice Centre of Digital and Public Humanities del Dipartimento di Studi Umanistici dell'Università Ca' Foscari di Venezia (VeDPH) ed il CNR-ILC, che ha sede all'interno dell'Area della Ricerca del CNR di Pisa[63]. La Zotero Library di DiPText-KC[64] è integrata nella sezione Bibliographic Resources del sito Web di DiPText-KC[65].

La presenza di risorse ben organizzate in repository centralizzati può costituire un vantaggio per gli studiosi dediti ad attività di ricerca. Ciò, infatti, consente di ridurre notevolmente il tempo impiegato per il reperimento di documentazione sul Web. Inoltre, come descritto nel capitolo precedente, il ruolo di un K-Centre consiste nel condividere la conoscenza e nel supportare la ricerca e una biblioteca digitale può rappresentare uno strumento adeguato a tale scopo.

Prima di creare la biblioteca vera è propria, è necessario sia reperire i materiali da inserire al suo interno in base agli argomenti che si intende trattare sia adottare una strategia che permetta di immagazzinarli. Una soluzione Open-Source ad hoc può essere individuata nell'applicazione Web Zotero, che permette di creare una biblioteca digitale con risorse scelte dall'utente (ad esempio, documenti in formato PDF, URL di siti Web ecc.) e di organizzare i contenuti trovati mediante l'utilizzo di cartelle, keyword e tag. Una volta terminato questo processo, il passo successivo consiste nell'integrazione della biblioteca digitale nell'ambito della tecnologia Web impiegata per la gestione dei contenuti. Nel caso del sito di DiPText-KC, tale tecnologia è costituita dall'applicazione Web WordPress.

La Zotero Library di DiPText-KC raccoglie alcune rilevanti voci bibliografiche relative ai domini delle Digital Humanities dedicati agli studi testuali. La raccolta si rivolge principalmente a studiosi e studenti di Filologia Digitale. I diversi passaggi che hanno portato alla realizzazione della prima release della biblioteca digitale del K-Centre sono descritti in dettaglio nelle sezioni seguenti.

3.1 Ricerca delle risorse e loro organizzazione in Zotero

Come si è detto, per immagazzinare le risorse si è scelto di fare ricorso a Zotero, uno strumento utilizzato per la gestione dei riferimenti bibliografici e dei materiali ad essi correlati (ad esempio, documenti in formato PDF). Le sue principali caratteristiche sono: la sua integrazione nell'ambito dei più diffusi Web browser ed editor di testo, la

[63] Il CNR-ILC ha anche un'Unità Organizzativa di Supporto (UOS) nell'Area della Ricerca del CNR di Genova ed un'Unità di Ricerca presso Terzi (URT) nella sede del VeDPH.
[64] Cfr. https://www.zotero.org/groups/4521720/diptext/library.
[65] Cfr. https://diptext-kc.clarin-it.it/knowledge/bibliographic-resources.
La prima release della Libreria DiPText-KC Zotero e la possibilità di sfogliarne gli elementi e le citazioni tramite il plugin Zotpress e gli appositi shortcode e widget nella sezione Bibliographic Resources del sito Web del K.Centre sono il risultato di un tirocinio curriculare svolto dal candidato presso il CNR-ILC sotto la guida della Dott.ssa Paola Baroni (Tutor Istituzionale) e del Dott. Angelo Mario Del Grosso (Tutor Scientifico) e con la consulenza del Dott. Federico Boschetti (Responsabile della URT veneziana del CNR-ILC e Referente di DiPText-KC sia per il CNR-ILC, rappresentato dalla Dott.ssa Monica Monachini, sia per il VeDPH, rappresentato dal Prof. Franz Fischer).

sincronizzazione online delle bibliografie e la generazione automatica di citazioni, note e bibliografie.

La prima operazione da fare consiste nel cercare le risorse da inserire nella biblioteca. Per farlo, è necessario individuare l'argomento principale che, nel caso della Zotero Library di DiPText-KC, è costituito dalle risorse relative al *Digital Scholarly Editing* e a tutte le sue fasi. È pertanto importante individuare siti Web dedicati alle pubblicazioni umanistiche proprio per estrarre i materiali d'interesse e, dopo la loro selezione, ripetere l'operazione di ricerca sulla bibliografia di ogni risorsa selezionata, riuscendo in tal modo a delineare un quadro esauriente dell'argomento d'indagine. Grazie all'estensione Zotero Connector di Google Chrome, è possibile salvare ciascuna risorsa direttamente sulla piattaforma Zotero scegliendo la cartella di destinazione.

Una volta terminata questa fase, la biblioteca conterrà tutte le risorse selezionate, visualizzabili nella sezione centrale dell'interfaccia. Nella sezione di sinistra, invece, l'interfaccia presenterà le cartelle relative all'utente (nella parte superiore), le cartelle relative al gruppo (nella parte intermedia) e i tag e le keywords per facilitare la ricerca dei materiali (nella parte inferiore). Nella sezione di destra l'interfaccia mostrerà infine le informazioni, i metadati, le note, i tag ed altri elementi bibliografici correlati alla risorsa individuata, all'utile scopo di permettere all'utente di farsi un'idea della risorsa senza doverla necessariamente aprire.

3.2 Configurazione di Zotero

In primo luogo, è opportuno scegliere quale versione di Zotero utilizzare: se la versione *Web-based*, o la versione *Desktop,* che richiede download e installazione.

In secondo luogo, è necessario installare nel browser che si intende usare il connettore Zotero adeguato, che permette di salvare i materiali d'interesse direttamente sulla piattaforma Zotero.

A questo punto è possibile creare la biblioteca digitale, configurando i vincoli di visualizzazione e di privacy e stabilendo se essa debba essere personale oppure condivisa con un determinato gruppo di utenti (da aggiungere dall'amministratore).

Una volta effettuata la configurazione iniziale ed attivato l'apposito plugin nel browser, è possibile salvare le risorse direttamente nella biblioteca ed ogni informazione relativa ad uno specifico argomento nell'apposita cartella previamente creata ad hoc.

3.3 Tagging tramite l'utilizzo della tassonomia TaDiRAH

Dopo aver inserito nella biblioteca le risorse d'interesse, è buona norma indicizzarle con keyword univoche e standardizzate al fine di facilitare la ricerca, rendendola chiara e veloce quanto più possibile.

La Zotero Library di DiPText-KC utilizza la tassonomia TaDiRAH (Taxonomy of Digital Research Activities in the Humanities). Tale tassonomia, sviluppata per essere

utilizzata nelle attività di ricerca digitale della comunità scientifica delle Scienze Umane, mira a raccogliere, strutturare e rendere facilmente individuabili ed accessibili informazioni relative a risorse, strumenti, metodi, progetti o articoli scientifici rilevanti per le Digital Humanities. La tassonomia TaDiRAH è articolata in diversi domini generali che corrispondono alle varie fasi del processo di ricerca. A ciascuno di questi domini è correlato un elenco chiuso di metodi che fanno riferimento ad attività nell'ambito di un dominio più ampio, il quale specifica cosa si sta facendo senza indicare in che modo lo si sta facendo. Sebbene si tratti di un elenco chiuso, esso può essere periodicamente revisionato.

La tassonomia dovrebbe essere integrata con due elenchi aperti separati: i) le **tecniche**, che forniscono informazioni più specifiche su come viene applicato un metodo. (ad esempio, *encoding*, *POS_tagging*, *technology_preservation*, *modelling* ecc.); ii) gli **oggetti**, che indicano l'oggetto o gli oggetti accademici a cui viene applicato un metodo o una tecnica. Questi elenchi aperti sono separati da obiettivi e metodi e una tecnica può essere associata a più di un metodo. (ad esempio, *bibliographic_listings, manuscript, methods* ecc.).

Le **attività** di ricerca sono generalmente applicate ad uno o più oggetti di ricerca. Un articolo sulla modellazione delle proprietà di un manoscritto viene contrassegnato con i tag "*modelling*" e "*manuscript*", mentre un normale editor di testo viene contrassegnato con i tag "*writing*", "*code*" e "*text*".

3.4 Il plugin Zotpress

Dopo l'organizzazione delle risorse mediante l'utilizzo della tassonomia TaDiRAH e una gestione appropriata della directory, va integrata la Zotero Library di DiPText-KC nel sito Web del K-Centre. A tale scopo si ricorre al plugin Zotpress di WordPress che, dopo una semplice configurazione, consente di importare la biblioteca digitale nel sito WordPress selezionato attraverso il gestore multipiattaforma Zotero.

3.4.1 Installazione e configurazione

Dopo averlo scaricato dalla sezione Plugin di WordPress ed averlo attivato, è possibile configurare Zotpress secondo le esigenze ritenute più opportune.

Per importare la biblioteca nel sito WordPress basta cliccare sulla voce "Zotpress" presente nella dashboard del sito e selezionare la voce "*Accounts*". La schermata che si apre presenta (nella parte superiore destra) il bottone "*Add Account*", premendo il quale si accede ad un form contenente i seguenti campi obbligatori: i) *Account Type*; ii) *API User ID*; iii) *Private Key*. L'*Account Type* va indicato in base al tipo di configurazione scelto per la Zotero Library (le opzioni possibili sono "*User*" e "*Group*"). L'*API User ID* per gli account *User* si trova in *Zotero settings > Keys*, mentre l'*API User ID* per gli account *Group* è contenuto nella URL della pagina del gruppo Zotero. La *Private Key*, una chiave univoca necessaria affinché Zotpress possa effettuare richieste da WordPress

a Zotero, può essere generata in *Zotero settings* > *Keys* scegliendo l'opzione "*Create new private key*".

Affinché la biblioteca sia visualizzabile, deve essere spuntata la voce "*Allow library access*". Per quanto riguarda i gruppi, inoltre, è necessario impostare su "Sola lettura" o su "Lettura/scrittura" le autorizzazioni (sia predefinite sia specifiche).

Cliccando sulla voce "*Browse*", la biblioteca viene quindi visualizzata. Per modificare le opzioni della biblioteca (come, ad esempio, l'account di default, lo stile della biblioteca o i widget da includervi), basta cliccare sulla voce "*Option*".

Per completare la procedura di installazione, è necessario creare una nuova pagina ed inserirvi uno "*shortcode*" che permetta di estrarre le categorie di dati che si intende far esplorare dal fruitore. È inoltre possibile personalizzare la configurazione aggiungendo funzioni quali, ad esempio, il tipo di ricerca e la possibilità di scaricare o di citare gli articoli individuati. Nella pagina "*Help*" dell'installazione di Zotpress sono disponibili il *basic shortcode [zotpressLib userid="000000"]* per realizzare una bibliografia *stand-alone* e un'ampia gamma di attributi per personalizzarla.

Lo *shortcode* inserito nella pagina Bibliographic Resources del sito Web di DiPText-KC è il seguente:

> *[zotpressLib userid="4521720" type="dropdown" dropdown="tags" sortby="date" order="desc" cite="yes" download="no" target="new"]*

Tale codice consente al fruitore di esplorare la biblioteca per argomento (tramite il menù dropdown di sinistra) o per tag (tramite il menù dropdown di destra). Gli elementi sono ordinati prima in ordine cronologico decrescente (dal più recente al meno recente) e poi in ordine alfabetico. Il download degli elementi non è consentito, mentre è possibile la loro citazione.

3.5 Vantaggi e svantaggi

La presenza di una biblioteca digitale all'interno di un *K-Centre* può costituire un vantaggio in termini di efficienza della ricerca. Il fatto che le risorse relative ad una disciplina specifica siano centralizzate in repository, raggruppate sulla base di una tassonomia scientifica ed organizzate in cartelle tematiche, infatti, garantisce non solo una notevole riduzione dei tempi della ricerca di articoli scientifici sul Web ma anche una maggiore persistenza - e, quindi, una maggiore portabilità - delle informazioni. Una maggiore flessibilità è inoltre assicurata dalla possibilità di aggiungere e rimuovere risorse, keyword e cartelle a seconda delle necessità individuate dall'admin.

L'unico svantaggio che la creazione di una biblioteca digitale di questo tipo presenta è costituito dal tempo richiesto dal reperimento iniziale delle risorse, che il realizzatore della bibliografia stand-alone deve necessariamente effettuare manualmente e con cura, impiegando una quantità di tempo che cresce all'aumentare del numero delle risorse che intende reperire.

Capitolo 4: L'Applicazione Web 'EasyScrape'

4.1 Introduzione

EasyScrape è un'applicazione Web concepita per estrarre dati bibliografici di articoli scientifici relativi al *Digital Scholarly Editing* da specifici archivi Web. Il processo di estrazione crea un insieme di riferimenti alle risorse con i rispettivi dati bibliografici quale il titolo, l'autore o l'anno di pubblicazione degli articoli. I dati così raccolti saranno organizzati e presentati all'utente per consentire una ricerca accurata e puntuale tra i vari articoli disponibili. Il codice sorgente dell'applicazione Web è disponibile nel repository GitHub[66] del CoPhiLab del CNR-ILC, che contiene tutte le funzioni e le strutture HTML e CSS realizzate dal candidato per la costruzione del front-end e del back-end di EasyScrape.

La tecnica implementata per estrarre le informazioni bibliografiche è quella del cosiddetto ***Web scraping***, una tecnica di acquisizione e recupero dati da sorgenti esterne. In particolare il Web scraping consiste nel recuperare un file in vari formati elaborabili dal calcolatore (HTML, XML, TXT ecc.) al fine di manipolazioni per esigenze più varie. Tra le varie applicazioni, questa tecnica viene spesso utilizzata per acquisire contenuti Web da fonti esterne presenti in rete su server differenti. Possiamo dividere il processo di Web scraping in 3 fasi:

1. **Acquisizione**: consiste nel selezionare e acquisire il contenuto Web d'interesse preparando opportunamente una richiesta HTTP/HTTPS ed interrogando la sorgente scelta.
2. **Manipolazione**: elaborare i dati acquisiti estraendo solo la parte di dati d'interesse manipolandola secondo le esigenze precipue.
3. **Produzione di un risultato**: utilizzare il contenuto acquisito costruendo una nuova risorsa che contenga i dati elaborati.

La bontà del metodo si basa sulla libera circolazione della conoscenza creando facilmente un ponte tra l'utente e i *provider* di risorse e di contenuti. Le norme legali sono particolarmente delicate quando governano il diritto d'autore e regolamentano gli utilizzi illeciti dei contenuti acquisiti.

Gli articoli interessati dal processo di *scraping* sono principalmente articoli scientifici, inerenti alle tecnologie sui linguaggi di *markup* adottate in ambito *Digital Scholarly Editing*. Le risorse fonte sono state estratte dai seguenti siti Web: *Balisage: The Markup Conference*; *Digital Scholarship in Humanities; Journal of the Text Encoding Initiative;* e *Google Scholar*.

Le tecnologie utilizzate sono quelle tipiche dello sviluppo Web, ovvero i linguaggi HTML, CSS, Javascript, PHP, SQL oltre all'ausilio di specifiche API (*Application*

[66] Cfr. https://github.com/CoPhi/easyscrape.

Programming Interfaces) utilizzate per scalare su un gran numero di richieste verso i servizi di interesse.

Le API sono elementi di una struttura intermedia, all'interno di un'architettura fondata sui microservizi, la quale consente a diverse applicazioni Web di comunicare tra di loro; dunque le API sono l'elemento base della comunicazione tra software. Il compito delle API è quello di ascoltare e verificare la validità delle comunicazioni oltre a fornire una risposta ad ogni richiesta valida in entrata; per fare ciò, un'applicazione Web con la funzione di client effettua un'operazione detta "chiamata API" con la quale interroga un server[67].

Il funzionamento delle principali funzioni di *scraping* implementate nell'applicazione rispecchia le tre fasi precedentemente indicate; l'acquisizione è la fase cruciale per ottenere il dato grezzo iniziale, da utilizzare prossimamente per le elaborazioni successive. In questa fase è solito riscontrare errori e anomalie, specialmente se il numero di richieste HTTP è elevato. È perciò necessario ricorrere all'utilizzo di servizi di API[68], denominate *Scrapestack* e *ScrapingBee*. La *query string* da spedire all'API in ascolto è formata dall'URL del sito bersaglio e dalla chiave univoca fornita dall'API. Successivamente mediante i metodi della libreria cURL di PHP è possibile avviare l'estrazione del contenuto Web.

Esistono vari strumenti per selezionare i dati da estrarre, per esempio mediante le procedure implementate dalla libreria *Gouttle* usando i selettori CSS o mediante la libreria *Simple HTML DOM*. EasyScrape utilizza la classe nativa `DOMDocument` di PHP per creare un pagina, la quale *"Rappresenta un intero documento HTML o XML; Esso rappresenta la radice dell'albero del documento"*[69]; sull'oggetto istanziato verranno poi applicati i metodi della classe `DomXpath`, così da estrarre il contenuto dagli elementi selezionati usando il linguaggio XPath che permette di individuare i nodi di qualsiasi insieme di elementi che costituiscono una pagina Web strutturata secondo il modello HTML DOM. Una volta individuati i selettori XPath è opportuno salvare i dati ottenuti dallo scrape in array associativi con appositi indici (ad esempio *author, title ,link, year*) i quali andranno ad indicare il tipo di dato da utilizzare nelle rielaborazioni future, come il salvataggio nel database o la visualizzazione e l'organizzazione in tabelle. La funzione di *scrape* che permette di estrarre e salvare massivamente tutti gli articoli da un sito Web, è costituita da 2 procedure distinte: la prima si occupa di recuperare gli URL delle varie pagine da cui estrarre gli articoli presenti nel sito riferito; la seconda procedura scorre l'array di link creato dalla funzione precedente ed estrae gli articoli da ogni risorsa. Nel caso in cui gli articoli si trovino in una singola pagina, è sufficiente la funzione apposita per i soli articoli.

[67] Cfr. la sezione "Cos'è una chiamata API" del sito Web di Openapi.
[68] *Scrapestack* (https://scrapestack.com) e *Scrapingbee* (https://www.scrapingbee.com) sono le API utilizzate nell'applicazione Web EasyScrape e consentono il rendering JavaScript, la personalizzazione delle intestazioni HTTP, l'utilizzo richieste POST/PUT e vari metodi di gestione del proxy.
[69] Cfr. la sezione "The DOMDocument class" del *PHP Manual* presente nel sito Web di PHP.net.

4.2 La funzione di *scrape*

Le funzioni di *scrape* mirano all'estrazione dei dati in modo automatico da pagine Web analizzando i vari collegamenti ipertestuali. Questa tecnica è molto usata nel marketing digitale, perché l'estrazione di dati permette di creare un prospetto statistico sui prodotti o sulle tendenze degli utenti. Le funzioni sviluppate su EasyScrape analizzano il DOM di un documento HTML estratto e successivamente selezionano e memorizzano le informazioni inerenti le singole pagine o gli elementi bibliografici degli articoli. Nelle sezioni successive verranno descritte nello specifico le fasi della procedura di *scrape*, citate in precedenza. Il codice della procedura implementata è riportato appendice A; in quel caso specifico la funzione estrae i link delle pagine dalla quale verranno estratti gli articoli.

4.2.1 Acquisizione della risorsa

La prima operazione del processo consiste nell'individuazione della risorsa. In questa fase vengono scelte le informazioni opportune ed il numero di pagine che si vogliono estrarre. Più elevato sarà il numero di pagine, quindi di risorse, più aumenterà il numero di richieste HTTP, e, di conseguenza, il tempo necessario ad ultimare le operazioni della procedura. Questo fattore causa in molti casi un reindirizzamento o un rifiuto dell'IP da parte del sito target, in modo da bloccare la procedura di *scraping*; fortunatamente le API utilizzate riescono a eludere l'ostacolo, grazie a chiamate HTTP riconosciute e accettate dal server target. Grazie a parametri determinati nella sessione cURL, lo *scraper* visita un URL specificato ed estrae una copia della pagina HTML dalla quale verranno estratti successivamente i dati inerenti le pagine e agli articoli.

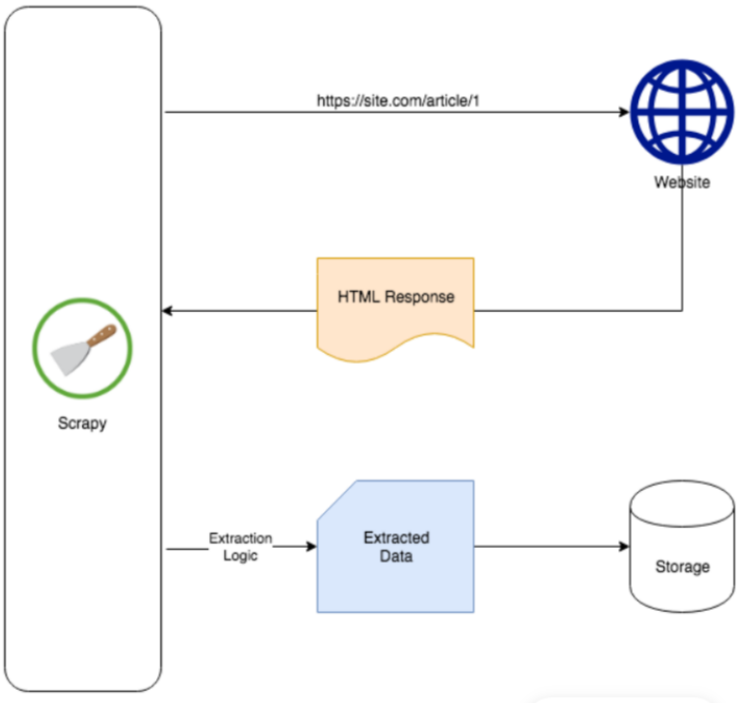

Immagine n° 1 - Schema sul funzionamento dello tecnica di *scrape*

EasyScrape è stato sviluppato per estrarre informazioni bibliografiche dai siti citati nell'introduzione e di seguito vedremo nel dettaglio l'acquisizione del contenuto da ogni sito Web. Da Balisage, vengono estratti tutti i proceedings e tutti gli articoli indicizzati per argomento (*topic)*, da DSH e JTEI tutti gli articoli divisi per anno, e infine i risultati delle prime 10 pagine della ricerca *Digital Scholarly Editing* da Google Scholar. Inoltre, per quest'ultimo contesto, l'amministratore dell'applicazione può aggiungere altre pagine a sua discrezione, inserendo semplicemente il link della pagina da cui estrarre le informazioni. Per portare a termine il processo di scraping da Balisage è stato sufficiente utilizzare i metodi della libreria *cURL*, mentre per tutti gli altri archivi è stato necessario impostare una specifica *query string* con l'ausilio dei servizi API sopracitati.

L'inizializzazione della funzione[70] inizia con il settaggio della *query string*, grazie al metodo integrato di PHP `http_build_query()`, questo permette di generare una *query string* che modifica l'URL della pagina *target*, utilizzando come parametro un array formato dall'*access key* dell'API e dalla variabile `$archiveURL`, una stringa che rappresenta il link della pagina di interesse.

Sprintf è un metodo nativo di PHP che permette di unire la *query string* con l'URL dell'API per ottenere il formato della query adatto al completamento della richiesta HTTP per poi effettuare la chiamata. Adesso grazie all'estensione cURL, istanziata nella variabile `$ch` con il metodo di PHP `curl_init()`, è possibile avviare questo tipo di sessione e inviare la richiesta all'API per attendere il risultato dal server. Con il metodo `curl_setopt()`, viene configurata la sessione cURL, le costanti che specificano i parametri della sessione e infine l'URL creato precedentemente. Le costanti usate in queste caso sono `CURLOPT_URL` e `CURLOPT_RETURNTRANSFER`, la prima consente il settaggio dell'URL per il trasferimento dei dati, la seconda invece, se settata al valore *true*, restituisce il risultato del trasferimento come una stringa e viene salvata nella variabile `$html` mediante il seguente metodo `$html=curl_exec($ch)`. Dopo questo primo passaggio vengono visualizzate le pagine HTML estratte contenenti gli articoli desiderati con i rispettivi link, dunque è possibile proseguire verso la selezione delle informazioni. Il codice sottostante è racchiuso in funzioni PHP, che hanno lo scopo di estrarre i link di tutte le pagine di ogni sito Web.

Prima di procedere con la fase successiva, resta da istanziare gli oggetti `DOMDocument` e `DomXPath` delle rispettiva classi[71] native di PHP. La prima classe viene utilizzata per creare l'albero DOM della pagina HTML estratta sotto forma di stringa utilizzando il metodo `loadHTML($html)`, il quale ha come parametro la pagina selezionata . Il secondo oggetto, quello della classe `DomXPath`, sarà utilizzato per inizializzare la

[70] Inizializzazione query string e cURL, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/includes/DSHscrape.php#L183, righe 183-197.
[71] Inizializzazione delle classi DOMDocument e DomXPath, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/includes/DSHscrape.php#L199.

ricerca XPath sul DOM. Dopodichè sarà possibile procedere con la manipolazione dei risultati.

4.2.2 Manipolazione

Dopo aver ottenuto la pagina Web, la fase successiva che la procedura di *scrape* esegue è la selezione dei dati. Questa fase è stata implementata con il supporto della classe `DomXPath` di PHP.

Lo schema dati scelto da EasyScrape definisce principalmente informazioni come l'autore, il titolo, il link e l'anno di tutti gli articoli presenti nella pagina attualmente selezionata. Spesso però risulta necessaria una ulteriore fase di ripulitura dei dati, come ad esempio la rimozione di caratteri speciali come le le virgolette (“”) o l'aggiunta di eventuali dati quando si presenta una lacuna, come nel caso in cui una pubblicazione è sprovvista di autore. È possibile individuare questo problema confrontando il numero dei risultati ottenuti in ogni indice dell'array associativo `$results`, il quale conterrà gli elementi estratti; nella sezione 4.2.3.2 si mostra come è possibile raffinare i dati grezzi ottenuti dal processo di *scrape*. Il codice sottostante riporta la procedura di estrazione dei dati realizzata con la classe `DomXPath`, la quale fornisce il metodo `query()` per cercare specifici elementi all'interno del DOM; in caso di esito positivo, vengono salvati gli elementi come oggetti con un indice specifico dell'array associativo `$results`, in base all'informazione estratta.

Successivamente viene preparato un nuovo array `$arr` al fine di memorizzare i vari elementi dell'articolo recuperando solo i dati necessari presenti negli oggetti salvati in `$results`.

```
$results = [];
$results['year']=$xpath->query('//h1[@id="publiTitle"]');
$results['author']=$xpath->query('//ul[@class="summary"]//li//div[@class
="author"]');
$results['title']=$xpath->query('//ul[@class="summary"]//li//div[@class=
"title"]//a');
$results['link']=$xpath->query('//ul[@class="summary"]//li//div[@class="
title"]//a//@href');

$arr = [];
```

Dopo aver osservato e compreso quali sono le informazioni da selezionare,è possibile accedere al codice sorgente della pagina per stabilire i tag e gli attributi da considerare nell'espressione XPath.

I selettori come `//h1[@id="publiTitle"]` permettono di estrarre il contenuto del tag specificando la classe o l'id per una maggiore precisione. In questo caso specifico si seleziona l'anno, che viene estratto direttamente dal titolo della pagina, perciò avrà un solo oggetto all'interno del sottoarray `$results['year']`, ma in altri casi dovrà essere allineato con il numero degli oggetti presenti negli altri sottoarray.

```
for($x=0; $x < $results['title']->length;$x++){
        $author = $results['author']->item($x)->textContent;
        $title = $results['title']->item($x)->textContent;
        $link = $results['link']->item($x)->textContent;
        $year = $results['year']->item(0)->textContent;
[...]
       $arr[$x]['title']   =  trim($title);
       $arr[$x]['author']  =  trim($author);
       $arr[$x]['link']    =  $jTeiCompl.trim($link);
       $arr[$x]['year']    =  $final_year;
       }
return $arr;
}
```

Listato n° 1 - Iterazione dell'array $result e salvataggio nell'array $arr

I selettori XPath e la sua sintassi sono illustrati nell'apposita sezione (4.2.3.1). Il codice del *listato n° 1* indica come estrarre informazioni dall'array `$results` utilizzando il metodo `item()` della classe `DOMDocument` e il successivo salvataggio in un altro array. Il codice mostrato nel *listato n° 1*, riguarda il riordinamento dei dati estratti con successivo inserimento nell'array `$arr`. Per prima cosa grazie al metodo `item($x)->textContent` insieme alla sua proprietà (`textContent`), è possibile recuperare il contenuto testuale del nodo e quindi memorizzarlo come stringa nelle rispettive variabili, l'indice `$x` sarà quindi fondamentale per riunire le informazioni di ogni articolo sotto un unico indice.

Il codice precedentemente presentato, mostra un'interruzione "[...]"; la parte di codice omessa sarà descritta nello specifico nella sezione 4.2.3.2, che tratta il raffinamento dei dati. Infatti, questi dati possono essere stringhe di varia natura, risulta quindi opportuno agire a seconda delle varie necessità e delle differenti casistiche. Tuttavia, analizzando le pagine di un sito è possibile notare che alcune di queste presentano una struttura diversa dalle altre pagine, ad esempio il titolo o l'anno potrebbero avere una collocazione diversa da pagina a pagina, rendendo così necessaria la previsione di tali casi per prevenire anomalie o errori che si potrebbero verificare durante l'elaborazione, quindi risulta necessaria la comprensione della struttura delle pagine di interesse. Oltre ad una collocazione diversa delle informazioni, può capitare anche di trovare una ripetizione o di trovare un elemento vuoto, per questo è opportuno agire con metodi che operano sulle stringhe tra cui `str_replace()`,`explode()` oppure `preg_replace()`, per attuare una corretta fase di *data cleaning*. Come si può notare in questo caso, le variabili `$jTeiCompl.trim($link)` e `$final_year` sono diverse da quelle definite all'inizio del ciclo *for*. Ad esempio, il link estratto viene concatenato con una variabile nella quale è presente l'indirizzo complementare[72] con la specificazione del protocollo utilizzato e del nome di dominio del sito dalla quale si estrae l' articolo visto che nel link recuperato dall'articolo generalmente non è presente

[72] Indirizzo complementare:
`$jTeiCompl='https://journals.openedition.org/jtei/.`

questa parte di URL. Nel caso dell'anno è necessario eseguire alcuni controlli ed eseguire correzioni in base alla casistica.

4.2.2.1 XPath

Vista la struttura gerarchica ad albero ordinato di una pagina HTML, XPath può essere utilizzato insieme per selezionarne i nodi, esattamente come se fosse scritto in linguaggio XML. XPath si basa sulla creazione di espressioni per selezionare i nodi dell'albero. Le espressioni utilizzate nella funzione di *scrape* hanno una struttura simile all'espressione seguente:

```
//ul[@class="summary"]//li//div[@class="title"]//a//@href
```

È possibile notare i tag HTML seguiti dal nome della classe per specificare la sezione, nello specifico, questa espressione recupera i link di un titolo nel seguente modo: grazie ai caratteri //, l'espressione `//ul[@class="summary"]//li` seleziona tutti gli elementi `<li>`, ovvero i nodi discendenti del tag `<ul>` appartenenti alla classe *summary*, indipendentemente da dove gli elementi si trovano all'interno del documento, per recuperare successivamente all'interno di ogni tag `<li>` l'elemento div inerente al titolo `//div[@class="title"]`, dove è contenuto il link , quindi il tag `<a>` con il suo attributo href, selezionabile con il simbolo @ come da specifica per ogni altro attributo (`div[@class="title"]//a//@href`). Quindi, l'estrazione indaga sempre più in profondità i nodi discendenti di un certo tag, fino a trovare l'informazione specificata nell'espressione XPath.

Quindi l'estrazione indaga sempre più in profondità i nodi discendenti di un certo tag, per cercare un'informazione specifica.

4.2.2.1.1 Metodo alternativo: selettori CSS

Oltre al linguaggio XPath è possibile utilizzare i selettori CSS per selezionare il contenuto HTML da estrarre. Un selettore CSS, come per XPath, è un'espressione che rappresenta un pattern di nodi all'interno dell'albero e ne restituisce il contenuto. Per selezionare ogni `<div>` presente nel codice sorgente, sarà sufficiente digitare nella *query* il nome del tag div, mentre nel linguaggio XPath verrà utilizzata l'espressione `//div`.

Le espressioni CSS sono costituite da selettori e sono gli stessi utilizzati per definire le regole di stile, i quali possono indicare un tag specifico, gli attributi degli elementi come classe o id. Questi selettori vengono utilizzati principalmente insieme alla libreria *Goutte* di PHP o *simple HTML DOM*. *Goutte* mette a disposizione alcuni metodi e classi per facilitare lo scraping; ad esempio fornendo l'oggetto *`$crawler`* nella quale cercare o il metodo `filter()`.

```
$crawler=$client->request('GET','https://www.imdb.com/search/name/?birth
_monthday=12-10');

$crawler->filter('div.lister-list h3 > a')->each(function ($node)
```

4.2.2.2 Sistemazione e correzione dei dati

Nel codice mostrato nel *listato n° 1*, è presente un'interruzione che sarà completata nella sezione corrente con il *listato n° 2*. Questa sezione riguarda il controllo sui dati prima del salvataggio nell'array `$arr`. Molte volte i dati recuperati presentano lacune, imperfezioni e caratteri speciali da eliminare rendendo quindi necessario l'utilizzo di metodi che operino sulle stringhe e altri che agiscono sulla dimensione dell'array per migliorare la qualità finale del dato. Di seguito saranno descritti e mostrati alcuni esempi specifici di fenomeni che si possono verificare.

Nel primo caso analizzato è presente una lacuna sul campo autore per vari articoli estratti dal sito DSH. Quando si presentano alcune *keywords* specifiche nel titolo si nota, nella pagina Web d'interesse, l'assenza di un autore nella sezione dedicata al raggruppamento degli articoli.

```
for($x=0; $x < $results['authors'] ->length;$x++){
     $auth_arr[$x] = $results['authors']->item($x)->textContent;
}
for($x=0; $x < $resultsVolume['title'] ->length;$x++){
   $txt = $resultsVolume['title']->item($x)->textContent;
   if( strpos($txt,'Erratum') !== false || strpos($txt,'Corrigendum')
!== false || (strpos($txt,'Introduction') !== false && strlen($txt) ==
12 ) ){
      $pos = $x;
      $val = 'Nessun risutato';
      $auth_arr = array_merge(array_slice($auth_arr, 0, $pos),
array($val), array_slice($auth_arr, $pos));}

}
```

Listato n° 2 - Sistemazione dell'array $results['authors'] in presenza di una lacuna

In questo caso si crea un array con tutti i risultati del sottoarray degli autori per renderlo di lunghezza uguale agli altri, effettuando dei controlli per aggiungere il campo `$val` nell'array `$auth_arr`, nella cella opportuna, quando non vengono trovati gli autori.

Dopo aver popolato `$auth_arr` e dopo aver individuato gli articoli che causano l'errore è possibile controllare se il titolo contiene la *keyword* stabilita grazie al metodo `strpos()`, il quale restituisce *true* se trova la parola all'interno della stringa. Inoltre grazie al metodo `array_merge()` combinato con `array_slice()`, presenti di default nel linguaggio php, è possibile creare uno spazio all'interno dell'array quando si verifica la condizione presente nel secondo ciclo *for* del *listato n° 2*. Il primo unisce uno o più array mentre il secondo rimuove una parte dell'array e la sostituisce.

Il codice presente nell'appendice B vuole risolvere una problematica sull'anno, visto che all'interno del titolo l'anno può trovarsi in posizioni diverse, comportando delle mancanze durante la selezione. Si individuano quindi le diverse casistiche controllando, grazie al metodo `strpos()`, la presenza delle *keywords* presenti nei vari casi. Nel primo caso, applicando due volte il metodo `explode()` sarà possibile avere il titolo, selezionando solo i caratteri numerici dalla stringa contenuta nell'array `$newstr`

grazie al metodo `preg_replace()`. Purtroppo alcune volte l'anno può essere ripetuto 2 volte, quindi è necessario eseguire un controllo anche sulla lunghezza della stringa per poi dividerla e selezionare la stringa corretta.

Il metodo `preg_replace()` fa parte di una famiglia di metodi, nativi di PHP, che consente l'utilizzo di espressioni regolari, infatti questo comando esegue una ricerca secondo l'espressione regolare stabilita e sostituisce il contenuto. Un'altro metodo simile è `preg_match()`, questo trova tutte le occorrenze di ciò che si sta cercando nell'espressione regolare. Per quanto riguarda le stringhe, in questo caso specifico, vengono usati i seguenti metodi integrati nel linguaggio PHP:

- `strpos()`: trova la posizione della prima occorrenza di una sottostringa in una stringa;
- `str_replace()`: sostituisce tutte le occorrenze della stringa di ricerca con la stringa di sostituzione;
- `explode()`: divide una stringa quando trova un certo carattere come separatore e i due pezzi di stringa vengono salvati in un array;
- `substr()`: restituisce parte di una stringa.

Un'altro problema frequente riguarda l'espressione dei link, questi infatti spesso sono URL relativi, del tipo *../dsh/advance-article/doi/10.1093/llc/fqac039/6659069.* Per rendere funzionante il link, è stato necessario aggiungere https://academic.oup.com/ per ottenere un link in forma assoluta .

4.2.3 Produzione del risultato

Una volta conclusa la manipolazione dei dati, il risultato finale sarà contenuto in un array con le informazioni riguardanti il titolo, l'autore, l'anno, il link della risorsa ed eventualmente altre informazioni di corredo come il topic oppure il numero di volume o issue di ogni articolo analizzato. Inoltre dopo il raffinamento dei dati è stato possibile implementare delle funzioni che sfruttano i risultati dello scrape. Come ad esempio le funzioni denominate `getBalisageTopic()`, `add_all_Issues()` o `delete_advanced_articles()`, le quali serviranno per l'interazione con il database.

Il database utilizzato da EasyScrape, chiamato *scrapedb*, è formato da 4 tabelle: la prima, denominata *articles*, contiene tutti gli articoli estratti con una struttura uniforme, come citato in precedenza, lasciando eventualmente vuoti (*null*) i campi per cui non viene riscontrato nessun valore associabile durante il processo di scraping. Per individuare il sito di appartenenza di un determinato articolo, si ricorre alla colonna *art_table_id*, la quale funge da *foreign-key* tra la tabella degli articoli *articles* e la tabella *tab_sites* dove si trovano le informazioni dei siti sorgente. Inoltre, sono presenti altre due tabelle: *users* utilizzata per inserire i dati degli utenti admin e *logs* che permette di tenere traccia della data dei vari inserimenti effettuati per ciascun sito.

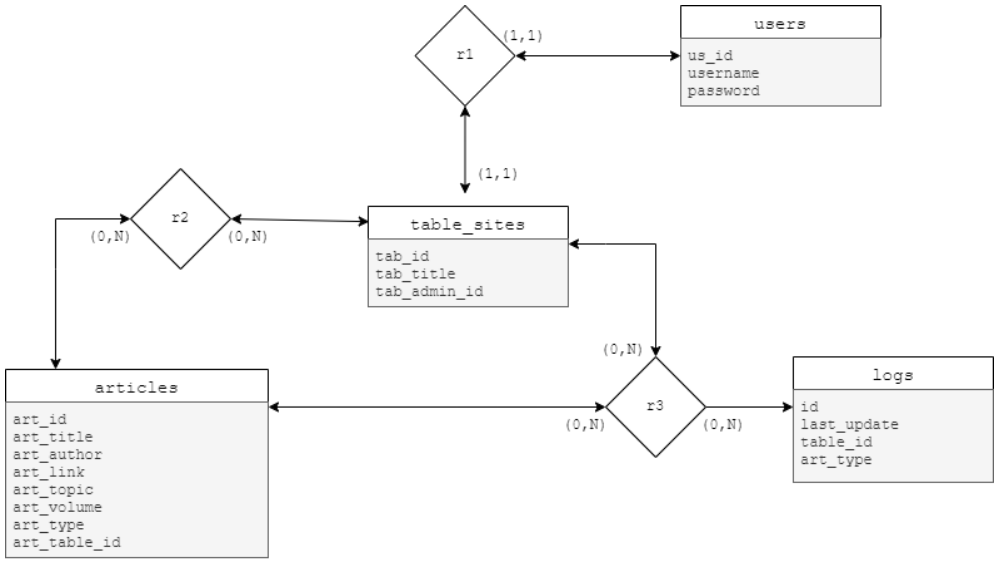

Immagine n° 2 - Diagramma E-R

I dati vengono inseriti nel database grazie alle funzioni che hanno il prefisso *add* (appendice C) e sono utilizzabili solo dall'utente admin nell'apposita area di *backend* dedicata. Prima di inserire effettivamente nel db i dati estratti, le funzioni di *add* richiamano al loro interno delle funzioni per effettuare lo *scrape*, come ad esempio la funzione `DOSCRAPEOFALLISSUE()`, la quale richiama a sua volta 2 funzioni di *scrape* differenti: la prima funzione ha il compito di istanziare un array contenente i link di tutte le pagine di un sito specifico, volte all'estrazione dei dati, un esempio di questa prima funzione è visibile nell'appendice A. La seconda utilizza come parametro l'array creato dalla prima funzione e ha il compito di estrarre i singoli articoli da ogni pagina, iterando l'array passato a tale funzione come parametro.

```
function DOSCRAPEOFALLISSUE($all_issues_archive){
      $issuesYear_arr = scrapeIssuesYear($all_issues_archive);
      $total_Issues = scrapeAllIssues_fromVolume($issuesYear_arr);
      return $total_Issues;

}
```

In questo caso è la funzione `scrapeIssuesYear($all_issues_archive)` a costruire un array con i link di tutte le pagine da analizzare, tale funzione utilizza un URL come parametro per cominciare l'estrazione dei link. Iterando l'array ottenuto da questa prima funzione è possibile estrarre tutti gli articoli presenti nei vari volumi, ricorrendo alla funzione `scrapeAllIssues_fromVolume($issuesYear_arr)`. Questa seconda funzione utilizza l'array con i link di ciascun volume per estrarre tutte le singole pagine dei volumi dove sono contenuti gli articoli. Dunque, invoca la funzione `scrapeVolume($linksVolume)` per ottenere un array chiamato `$volumes` contenente le singole pagine di ciascun volume. Infine, per ottenere gli elementi bibliografici di ogni articolo, viene utilizzata un'altra funzione di *scrape* denominata `ScrapeOneIssue($volumeIssue)`, la quale avrà un codice molto simile a quello

presentato nell'appendice A, ma con selettori XPath diversi da quelli visibili in tale sezione.

```
function scrapeAllIssues_fromVolume($issuesYear_arr){
  $volumes = [];
  foreach($issuesYear_arr as $k => $linksVolume){
          $volumes[] = scrapeVolume($linksVolume);
  }
  foreach($volumes as $key => $volumeLinkArr){
       foreach($volumeLinkArr as $index => $volumeIssue){
             $total_issues[] = ScrapeOneIssue($volumeIssue);}
  }
 return $total_issues;

}
```

Listato n° 3 - Funzione di *scrape* - Estrazione articoli

La funzione `DOSCRAPEOFALLISSUE()` risulta quindi essere il primo passo per l'estrazione dei dati e per la memorizzazione nel database (Appendice B), poiché recupera tutti gli articoli da salvare. Una volta ottenuto l'array contenente i risultati dello scrape è opportuno riorganizzare l'array, regolarizzando i suoi indici dato che gli articoli vengono raggruppati in sottoarray in base al topic, anno o volume di appartenenza, infatti questo array possiede un primo livello con l'indice del volume e il secondo indicherà i singoli articoli. La modellazione viene svolta con l'utilizzo di cicli *foreach*, dopodichè si preparano le variabili con i valori da salvare all'interno di un ciclo, il quale andrà ad eseguire la *query* specificata: ad ogni iterazione si salva il contenuto estratto dalle pagine all'interno del database.

Prima di questo passaggio, la funzione effettua un controllo sull'esistenza o meno dell'articolo all'interno del database. Una volta terminato il ciclo viene inserita la data e l'identificativo del sito nella tabella *logs,* per monitorare la data degli ultimi inserimenti degli articoli di un certo tipo. Questa funzione richiede un tempo non trascurabile prima che sia completata: il problema principale riguarda il numero di elementi da estrarre, che diventa elevato quando si cicla un array contenente gli URL di molte pagine.

La funzione che aggiunge gli articoli da Google Scholar, oltre agli articoli presenti di default sul *Digital Scholarly Editing,* permette all'utente amministratore di inserire l'URL di una pagina nell'apposito form del *backend* ed estrarre gli articoli della pagina scelta, rendendo quindi personalizzabile la scelta degli articoli reperibili da Google Scholar.

Un'altra funzione presente nell'area admin riguarda la cancellazione[73] di un certo numero di articoli in base al sito di appartenenza o al tipo di articolo, ed è visibile nel *listato n° 4*.

[73] Funzione *delete*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/includes/functions.php#L370.

```
function delete_all_Issues(){
   global $conn;
   $delete_sql = "DELETE FROM articles where art_type ='IS' AND
art_table_id = 2";
   $result = mysqli_query($conn, $delete_sql);

}
```

Listato n° 4 - Funzione *delete*

Questa operazione viene svolta cliccando semplicemente un bottone con un'etichetta specifica. Ovviamente questa possibilità viene riservata all'amministratore, in quanto gestore dell'applicazione e degli articoli. Per fare ciò si indica nella condizione *where* la categoria dell'articolo e l'indice che rappresenta il sito di provenienza, situato nella tabella *tab_sites*.

Le funzioni con il prefisso *get*[74], recuperano le informazioni all'interno del database e vengono utilizzate indirettamente anche dall'utente fruitore. Queste servono principalmente per la visualizzazione dei dati presenti all'interno delle tabelle per la ricerca degli articoli memorizzati. Come per le funzioni di *delete*, si nota la stessa struttura con i campi definiti in base al sito o al tipo di articolo.

Le tabelle memorizzano i seguenti dati : autore, titolo, anno e, se presenti, volume e topic. Questi dati vengono recuperati dal database, dalle funzioni *get,* e successivamente organizzati all'interno di tabelle HTML. I tag `<td>` contengono quindi le singole informazioni testuali, ovvero i risultati dello scrape, ad esempio, un tag di questo tipo contiene la stringa con il valore del titolo, inserita a sua volta nel tag `<a>` il quale avrà come attributo *href* il link della risorsa recuperata con lo *scraping*, in modo tale da rendere cliccabile il link sulla tabella. La creazione di una tabella, realizzata nel suddetto modo, contenente tutti i proceedings estratti Balisage, è visibile nell'appendice C.

Per la strutturazione e visualizzazione delle tabelle si ricorre al Plugin "*datatables*", una libreria JQuery di SpryMedia Ltd che permette la realizzazione di tabelle dinamiche e interattive. Queste tabelle sono personalizzabili grazie ad una moltitudine di opzioni per l'impaginazione, il filtraggio e l'ordinamento dei dati, garantendo un alto livello di interattività. L'inizializzazione della tabella viene fatta utilizzando il metodo `.DataTable()`[75] e i suoi parametri garantiscono una personalizzazione in funzione delle esigenze.

In questo caso specifico i primi 3 parametri indicano la risposta che dovrà essere data dalla tabella in caso di mancanza dei dati (*'sInfoEmpty','sEmptyTable','sZeroRecords'*), `'searching':true`, abilita la ricerca e mostra la casella di testo allo scopo definita,

[74] Funzione *get*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/includes/functions.php#L266.
[75] Configurazione *datatable*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/includes/datatable_config.php#L7.

`'paging':true` abilita l'impaginazione della tabella mentre `'pageLength':10`, indica il numero di records visualizzabili contemporaneamente; *`'ordering':true`* consente l'ordinamento degli articoli visualizzati, scegliendo per quale campo ordinare. E infine *`'searchHighlight': true`* evidenzia i caratteri del risultato della ricerca, ad esempio cercando la parola XML, questa verrà evidenziata in tutti gli articoli filtrati. Adesso è possibile visualizzare, cercare e ordinare le informazioni recuperate con la tecnica di scraping per fornire quindi all'utente uno strumento estremamente portatile, contenente svariate risorse sul mondo del markup per le edizioni digitali scientifiche.

Nell'applicazione Web, sopra le tabelle appena descritte, è presente un form che rappresenta un filtro per attuare una ricerca avanzata. Questa ricerca consente di selezionare i dati in base a delle condizioni scelte dell'utente, creando così una ricerca aggregata per autore, titolo, anno o per un certo intervallo temporale e se presenti anche per i volumi e per i topic. La funzione utilizzata dal *form* per effettuare la ricerca avanzata è presente nell'appendice F. La funzione utilizza come parametro i risultati del form di ricerca contenuti nella variabile `$_POST`, se le variabili non sono vuote viene creata una condizione aggiuntiva nella clausola *where,* utilizzando gli operatori SQL *LIKE* o *AND* seguito da una specifica condizione SQL; nel caso in cui fosse abilitato l'inserimento di più valori (come nel caso dell'anno o del volume) per lo stesso tipo di informazione si crea una stringa concatenata contente più operatori *AND* per poi inserirla successivamente nella clausola *where* insieme alle altre stringhe contenenti le condizioni generate dai valori ottenuti dal *form*, per poi eseguire la *query* e ottenere così il risultato desiderato.

4.3 Ulteriori funzioni e componenti utilizzati

L'intera struttura HTML è stata realizzata utilizzando la versione 4 di *Bootstrap*, un noto *framework Open Source* utilizzato per lo sviluppo di siti Web *responsive*. Nello specifico, il *framework* offre file CSS e Javascript che consentono l'utilizzo di classi per organizzare i vari contenuti. La struttura di base utilizza 12 colonne per organizzare i contenuti. Ad esempio `class="col-sm-2"` indica l'utilizzo di una colonna di grandezza 2 e specifica, grazie all'indicatore "sm", la dimensione dello schermo nella quale si applicheranno determinate regole di visualizzazione. Più classi concatenate (`class="col-xl-6 col-md-4 col-sm-2"`) specificano le diverse regole CSS da applicare in base alla grandezza del dispositivo, questo fattore è di fondamentale importanza per creare un sito Web in ottica *responsive*. Sempre di *Bootstrap* è il plugin che ottimizza l'utilizzo e la visualizzazione della selezione multipla del tag `<select>` fornendo una classe chiamata *selectpicker*.

Il codice che andrà a comporre il *select picker* conterrà un ciclo *foreach* per aggiungere tutte le opzioni contenute all'interno di un array, questo array avrà i valori restituiti dalle funzioni di *get* che recuperano solo gli anni, volumi o topic, dal database (come ad

esempio `getYearGS()` o `getBalisageTopics()`). Il *selectpicker*[76] sarà istanziato inserendo l'apposita classe nel suddetto attributo e utilizzando il metodo `$('.selectpicker').selectpicker()` sulla classe stessa.

Il codice Javascript presente nell'applicazione, è implementato principalmente con l'ausilio della libreria JQuery e l'utilizzo delle chiamate Ajax, per l'elaborazione dei dati senza dover ricaricare la pagina. Il codice presentato nell'appendice D riporta una chiamata Ajax che esegue una funzione di inserimento dei dati all'interno del database e allo stesso tempo attiva il loader, bloccando eventuali altre azioni sulla pagina. Alla fine dell'elaborazione, verrà nascosto il loader e verrà mostrato l'esito dell'elaborazione.

Al click di un bottone nell'area dell'amministratore (visibile nell'immagine n° 5 della sezione 4.3.2 Layout e logo), verrà recuperato il valore della tabella in questione, situato in un input di tipo *hidden,* insieme ad una stringa definita come *type* e saranno passati successivamente alla pagina *manage_articles.php*[77], dove sono presenti le chiamate alle funzioni PHP che consentono l'inserimento dei dati permettendo quindi la gestione del *back-end*.

Mentre i dati vengono elaborati da questo codice php, la chiamata Ajax mostrerà, grazie alla sintassi offerta da JQuery, il blocco dello schermo, nascosto all'avvio, dove all'interno è presente una gif di caricamento, il loader; il quale sarà successivamente chiuso una volta terminata l'elaborazione e verrà mostrata una finestra in modale, fornita dal *framework Bootstrap* 4 e aperta dalla funzione JS denominata *openModal()*.

4.3.1 Corsa critica

Le chiamate Ajax che invocano le funzioni di aggiunta dei dati richiedono un tempo variabile in base alla porzione di dati che si stanno elaborando; in linea teorica l'utente potrebbe cliccare su altri bottoni o link avviando quindi altri processi, portando ad una situazione di mutua attesa tra i vari processi. La mancata gestione di questo fattore comporta anomalie di vario tipo sull'inserimento dei dati, andando così a creare possibili errori in casi peggiori. E' opportuno gestire quella che viene definita corsa critica, ovvero una situazione in cui l'esecuzione di più processi dipende dall'ordine di esecuzione dei singoli processi, i quali accedono a risorse condivise.

La logica che EasyScrape adottata per controllare la gestione dei processi, riguarda un blocco totale del sistema che impedisce all'utente (e agli altri utenti) qualsiasi altra azione di modifica sul database. In questo modo l'utente dovrà aspettare il tempo necessario al completamento dell'elaborazione prima di poter fare qualsiasi altra azione sulla pagina corrente; così facendo sarà effettuato un inserimento per volta, senza far accavallare più processi tra di loro.

---

[76] Plugin *SelectPicker*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/balisage.php#L202.
[77] Pagina *manage_articles.php*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/admin/rpc/manage_articles.php.

Per creare questo blocco[78] è stata utilizzata una sezione `<div>` con le stesse dimensioni dell'intera pagina, al suo interno è stata inserita una gif per rappresentare il *loader* e il testo indicante lo stato dell'elaborazione.

Grazie alle regole CSS[79], nello specifico alla proprietà *display* con i suoi attributi *none* e *block,* è possibile nascondere o mostrare l'intera sezione di caricamento. Inoltre la proprietà *z-index* permette di specificare l'ordine di sovrapposizione degli elementi: un elemento con un valore *z-index* maggiore di un altro sarà sempre davanti ad un elemento con un valore *z-index* minore, creando in questo caso una sorta di "sipario" per il livello sottostante.

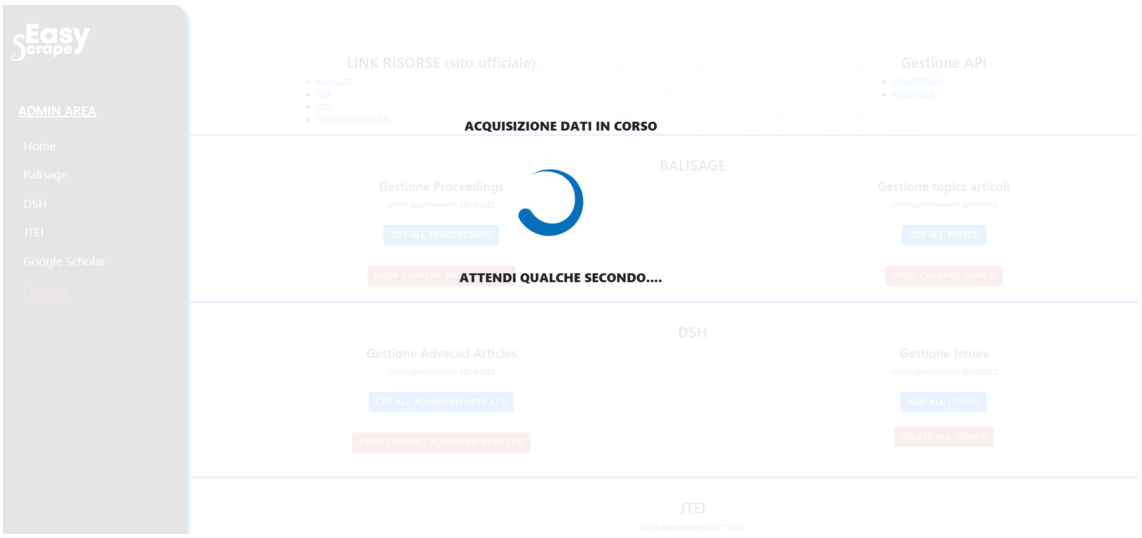

Immagine n° 3 - Blocco delle operazioni durante l'elaborazione della funzione di inserimento

4.3.2 Layout e logo

La struttura HTML del sito Web è stata realizzata con *Bootstrap* 4, ovvero un *framework* che aiuta lo sviluppatore nella realizzazione del *front-end* di un sito Web. *Bootstrap* fornisce anche CSS di default per personalizzare elementi come ad esempio la barra di navigazione, i bottoni, i blocchi strutturali (div) e il footer. I colori dominanti che sono stati utilizzati sono il bianco e il nero per dare un aspetto elegante e minimale all'applicazione web; per i dettagli invece è stato usato principalmente un azzurro accesso, indicato da *bootstrap* come *primary* (#007bff) e l'arancione su alcuni effetti *hover*.

[78] HTML riguardante il *loader*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/admin/admin.php#L322, righe 322-330.
[79] CSS riguardante il *loader*, https://github.com/CoPhi/easyscrape/blob/d397ac3ba2debf4609d233f7852f5888574708ee/progettoTesi/admin/admin.php#L107, righe 107-122.

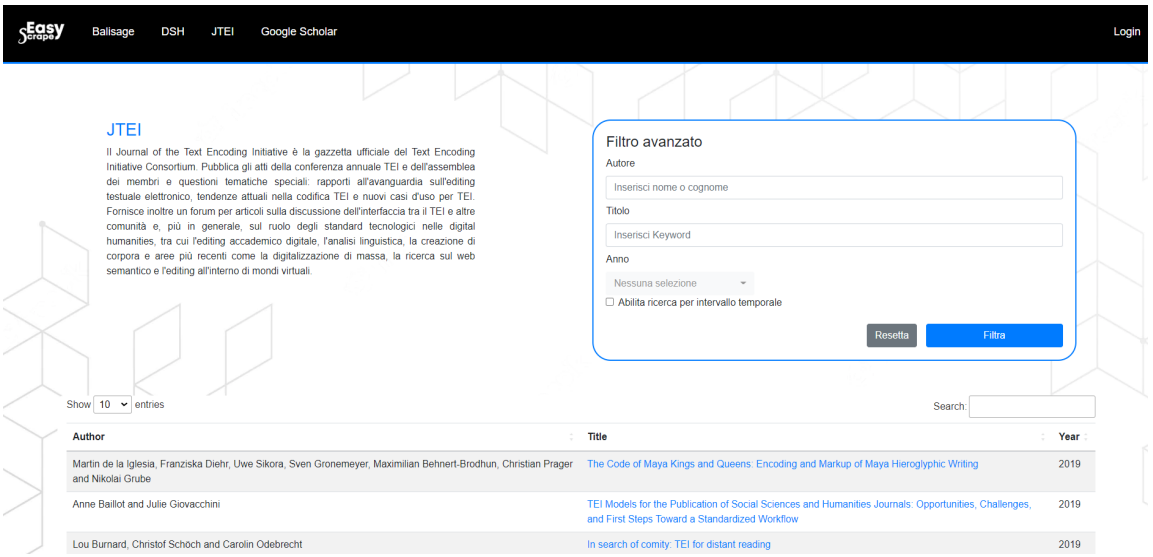

Immagine n° 4 - Front-end di EasyScrape

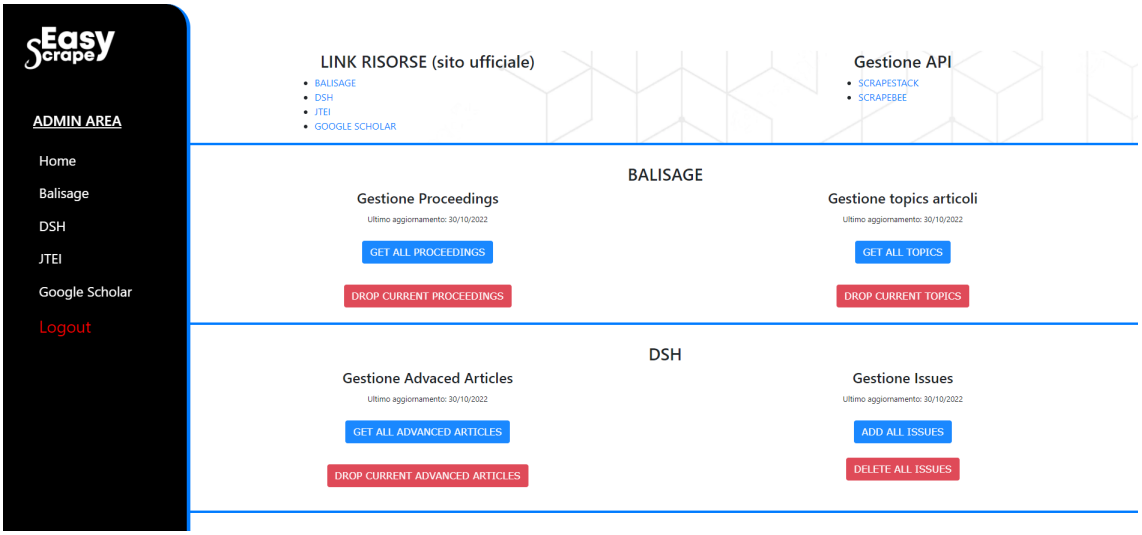

Immagine n° 5 - Back-end di EasyScrape

Per quanto riguarda il logo è stato realizzato con il software *Adobe Illustrator*; questo risulta essere una semplice scritta con un dettaglio particolare, ossia una "S" simile ad un amo da pesca, che vuol riprendere in un certo senso il concetto di "pescare nel Web". Questo dettaglio è stato realizzato vettorializzando l'immagine di un amo, la quale è stata successivamente unita al tracciato della "S".

4.4 Aspetti legali

EasyScrape non reperisce materiali per scopi economici. Intende, piuttosto, far circolare la conoscenza, salvaguardando la proprietà intellettuale e commerciale di ogni articolo raccolto. Fa quindi da interfaccia tra l'autore e l'utente finale, organizzando un notevole numero di link che indirizzano l'utente verso il repository in cui è depositato l'articolo.

La proprietà intellettuale si suddivide in proprietà industriale e diritto d'autore. Per garantire la proprietà industriale, la legislazione adotta vari tipi di privative industriali,

come brevetti o marchi. Il diritto d'autore protegge l'attività intellettuale, riconoscendo all'autore dell'opera tutti i diritti morali e patrimoniali. A tal proposito, l'Italia ha dato attuazione alla Direttiva 2001/29/CE del Parlamento Europeo e del Consiglio dell'UE del 22 maggio 2001 sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione[80] con il Decreto Legislativo n. 68 del 9 aprile 2003 - Attuazione della Direttiva 2001/29/CE sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione[81], dove così viene definita la protezione del diritto d'autore:

> "*Art. 1*
> *Sono protette ai sensi di questa legge le opere dell'ingegno di carattere creativo che appartengono alla letteratura, alla musica, alle arti figurative, all'architettura, al teatro ed alla cinematografia, qualunque ne sia il modo o la forma di espressione*."

Un sito Web rientra quindi nelle opere dell'ingegno di carattere creativo. In merito alla diffusione dei materiali, il suddetto decreto offre una panoramica sulle possibilità d'uso:

> "*Art. 65*
> *1. Gli articoli di attualità di carattere economico, politico o religioso, pubblicati nelle riviste o nei giornali, oppure radiodiffusi o messi a disposizione del pubblico, e gli altri materiali dello stesso carattere possono essere liberamente riprodotti o comunicati al pubblico in altre riviste o giornali, anche radiotelevisivi, se la riproduzione o l'utilizzazione non è stata espressamente riservata, purché si indichino la fonte da cui sono tratti, la data e il nome dell'autore, se riportato.*
> *2. La riproduzione o comunicazione al pubblico di opere o materiali protetti utilizzati in occasione di avvenimenti di attualità è consentita ai fini dell'esercizio del diritto di cronaca e nei limiti dello scopo informativo, sempre che si indichi, salvo caso di impossibilità, la fonte, incluso il nome dell'autore, se riportato."*

Con la Sentenza della Corte dell'UE (Quarta Sezione) del 13 febbraio 2014[82] espressa nella causa C-466/12, la giurisprudenza comunitaria sancisce che il rinvio con un link a contenuti di un altro sito Web, rispettando la proprietà intellettuale altrui, è lecito.

Tra le licenze *free content* più usate ci sono le ***Licenze Creative Commons*** (CC). Si tratta di strumenti gratuiti che garantiscono il diritto d'autore ed altri diritti connessi e consentono il riutilizzo, la condivisione ed eventuali modifiche in base a criteri stabiliti.

---

[80] Cfr. EUR-Lex - Direttiva 2001/29/CE del Parlamento Europeo e del Consiglio dell'Unione Europea del 22 maggio 2001 sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione.
[81] Cfr. Normattiva - Decreto Legislativo n. 68 del 9 aprile 2003 - Attuazione della Direttiva 2001/29/CE sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione.
[82] Cfr. Corte di Giustizia dell'Unione Europea (CURIA) - Sentenza della Corte (Quarta Sezione) del 13 febbraio 2014.

Le licenze CC in uso sono sei:

- **CC BY**: consente di distribuire, modificare, creare opere derivate dall'originale, anche a scopi commerciali, a condizione che venga riconosciuta una menzione di paternità adeguata, fornito un link alla licenza ed indicato se sono state effettuate delle modifiche;
- **CC BY-SA**: permette le stesse operazioni consentite dalla CC BY e consente di attribuire alla nuova opera la stessa licenza dell'originale, permettendo quindi l'uso commerciale di ogni opera derivata;
- **CC BY-ND**: permette le stesse operazioni consentite dalla CC BY ad esclusione della distribuzione di opere modificate, remixate o basate sull'opera soggetta a questa licenza;
- **CC BY-NC**: permette le stesse operazioni consentite dalla CC BY ma non a scopi commerciali e non impone di attribuire alla nuova opera la stessa licenza dell'originale;
- **CC BY-NC-SA**: permette le stesse operazioni consentite dalla CC BY ma non a scopi commerciali e impone di attribuire alla nuova opera la stessa licenza dell'originale;
- **CC BY-NC-ND**: consente solo di scaricare e condividere i lavori originali a condizione che non vengano modificati né utilizzati a scopi commerciali e sempre attribuendo la paternità dell'opera all'autore.

Il *quid* in più delle licenze CC è la 'struttura a tre livelli', per cui ogni licenza CC è un unico strumento giuridico ma può manifestarsi in tre forme differenti:

- ***Legal code***: costituisce la licenza vera e propria, cioè un documento dotato di valore legale in cui si disciplina la distribuzione dell'opera e l'applicazione della licenza;
- ***Commons deed***: non è una vera e propria licenza e non è dotato di valore legale, ma riassume nel modo più semplice possibile il contenuto della licenza;
- ***Digital code***: è costituito da una serie di metadati che rendono la licenza facilmente rintracciabile dai motori di ricerca e/o da macchine e strumenti automatici e ha lo scopo di identificare e catalogare automaticamente la licenza e le relative informazioni (attribuzione ecc.)[83].

Conclusione

Le infrastrutture di ricerca sono il cuore pulsante dello sviluppo e dell'innovazione della ricerca scientifica, in quanto affrontano le nuove sfide che scaturiscono dall'interazione, come nel caso del *Digital Scholarly Editing*, tra domini disciplinari diversi e tecnologie avanzate. Per raggiungere un'eccellenza scientifica riconosciuta a livello mondiale, il

[83] I metadati sono scritti in formato RDF grazie al *Creative Commons Rights Expression Language* (CC REL), una specifica tecnica che definisce come esprimere le informazioni della licenza in formato RDF e come integrare i metadati nell'opera. Tra le applicazioni pratiche del *Digital code* ci sono la possibilità di inserire le informazioni della licenza direttamente nel file e la possibilità di filtrare i risultati di una ricerca svolta su Internet o su un computer locale.

sostentamento delle infrastrutture di ricerca risulta fondamentale sotto ogni punto di vista. Per questo l'Unione Europa tende a valorizzarle finanziando progetti e iniziative comunitarie. L'efficienza e il coordinamento di queste infrastrutture è quindi la chiave per garantire un livello di qualità elevato alla ricerca e una visione sempre improntata al futuro.

Raccogliendo e organizzando opportunamente un gran numero di articoli e di risorse e supportando l'utente con varie metodologie volte alla ricerca di informazioni all'interno di una biblioteca digitale, le due soluzioni descritte nel terzo e nel quarto capitolo di questo elaborato offrono un contributo al miglioramento dell'efficienza della ricerca di articoli scientifici sul Web. Internet è il più grande serbatoio di contenuti mai esistito e la ricerca manuale di articoli scientifici validi al suo interno può essere lunga e dispendiosa, rendendo necessari successivi controlli sulla qualità del materiale reperito. È importante quindi creare biblioteche digitali tematiche per l'organizzazione di contenuti reperiti sul Web di cui si sia verificata la provenienza e l'alta qualità.

La prima soluzione adottata in questo lavoro propone l'integrazione di due applicazioni Web ormai consolidate a livello mondiale, dato l'ampio numero di siti realizzati con WordPress e di biblioteche realizzate con Zotero. Questa integrazione dà vita ad una biblioteca digitale semplice da realizzare e intuitiva da utilizzare. Inoltre, la ricerca dei contenuti è resa ancora più agevole grazie ad una tassonomia ben definita e all'organizzazione tematica delle cartelle. La seconda soluzione propone l'impiego della tecnica del *Web scraping*, grazie alla quale nel sito EasyScrape vengono analizzati ed estratti articoli presenti in quattro siti Web di interesse per le discipline umanistiche, per un totale di più di 1.000 articoli inerenti al *Digital Scholarly Editing*. Potenzialmente, esso potrebbe raccogliere ogni tipo di informazione o di risorsa scientifica: all'aumentare delle pagine, aumenterebbe infatti anche il numero degli articoli, trasformando così il sito in un grande serbatoio di informazioni globale, centralizzato e, soprattutto, ad accesso libero.

Nel primo caso si propone quindi una soluzione che utilizza strumenti Open Source facilmente configurabili ma che presenta una limitazione: la raccolta manuale degli articoli dal Web, infatti, richiede un grande dispendio in termini di tempo. Nel secondo caso si propone invece una soluzione che riduce notevolmente il tempo di raccolta delle informazioni, in quanto prevede di estrarle da "siti bersaglio" con un semplice click. Potenzialmente, la portata della sezione di *backend* gestita dall'admin potrebbe essere ampliata mediante l'aggiunta di bottoni che gestiscano l'inserimento e la cancellazione di articoli presenti anche in altri siti che non siano i quattro sopra citati, utilizzando le medesime funzioni di *scrape* e modificando soltanto la sintassi XPath per selezionare il nodo HTML che si vuole estrarre.

È pertanto possibile concludere affermando che un sito Web estendibile che includa, nel rispetto del diritto d'autore, collegamenti ipertestuali agli articoli scientifici presenti in altri siti può contribuire a reperire in modo efficace all'interno della Rete informazioni rilevanti per uno specifico dominio di conoscenza come il *Digital Scholarly Editing* e a facilitare così il lavoro degli studiosi che utilizzano le infrastrutture di ricerca.

Bibliografia

Articoli e documenti scientifici

Sahle, P. (2016), “2. What is a Scholarly Digital Edition?”. In Driscoll, M. J. and Pierazzo, E. (Eds.), *Digital scholarly editing: Theories and practices* (pp. 19–40), Digital Humanities Series, Vol. 4, Cambridge (UK): Open Book Publishers, http://dx.doi.org/10.11647/OBP.0095.

Burnard, L. (2014), *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*, Marseille (FR): OpenEdition Press, http://dx.doi.org/10.4000/books.oep.426.

Boschetti, F. and Del Grosso A. M. (2020), “L’annotazione di testi storico-letterari al tempo dei social media”. In *Italica Wratislaviensia* (Vol. 11, no. 1, pp. 65–99), Torun (PL): Wydawnictwo Adam Marszałek, http://dx.doi.org/10.15804/IW.2020.11.1.03.

Vagionakis I., Del Gratta R., Boschetti F., Baroni P., Del Grosso A. M., Mancinelli T. and Monachini M. (2021), “ ‘Cretan Institutional Inscriptions’ Meets CLARIN-IT”. In Monachini, M. and Eskevich M. (Eds.), *CLARIN Annual Conference Proceedings 2021* (pp. 48-53), Utrecht (NL): CLARIN ERIC, https://office.clarin.eu/v/CE-2021-1923-CLARIN2021_ConferenceProceedings.pdf.

Vagionakis, I., Del Gratta R., Boschetti F., Baroni P., Del Grosso A. M., Mancinelli T. and Monachini, M. (2022), “ ‘Cretan Institutional Inscriptions’ Meets CLARIN-IT”. In Monachini, M. and Eskevich M. (Eds.), *Selected Papers from the CLARIN Annual Conference 2021* (pp. 139-150), Linköping Electronic Conference Proceedings Series, Vol. 189, Linköping (SE): Linköping University Electronic Press, https://doi.org/10.3384/9789179294441.

Rossi L. C. (2003), “Finalità e metodi della filologia”, ICoN Italian Culture on the Net, Modulo 276, https://fdocumenti.com/document/finalita-e-metodi-della-filologia-unibgit-e-metodi-45-fase-interpretativa.html.

Piani, decreti, regolamenti, direttive e sentenze

Ministero dell'Università e della Ricerca - Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027, https://www.mur.gov.it/sites/default/files/2021-10/Decreto%20Ministeriale%20n.1082%20del%2010-09-2021%20-%20PNIR%202021%20-%202027.pdf (consultato in data 22/05/2022).

Ministero dell'Università e della Ricerca - Decreto Ministeriale n.1082 del 10-09-2021 - Adozione del Piano Nazionale Infrastrutture di Ricerca (PNIR) 2021-2027, https://www.mur.gov.it/sites/default/files/2021-10/Decreto%20Ministeriale%20n.1082%

20del%2010-09-2021%20-%20PNIR%202021%20-%202027.pdf (consultato in data 22/05/2022).

EUR-Lex - Regolamento (CE) n. 723/2009 del Consiglio dell'Unione Europea del 25 giugno 2009 relativo al quadro giuridico comunitario applicabile ad un consorzio per un'infrastruttura europea di ricerca (ERIC), https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32009R0723&from=LV (consultato in data 23/06/2022).

EUR-Lex - Direttiva 2001/29/CE del Parlamento Europeo e del Consiglio dell'Unione Europea del 22 maggio 2001 sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione, https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32001L0029&from=IT (consultata in data 16/09/2022).

Normattiva - Decreto Legislativo n. 68 del 9 aprile 2003 - Attuazione della Direttiva 2001/29/CE sull'armonizzazione di taluni aspetti del diritto d'autore e dei diritti connessi nella società dell'informazione, https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2003;068 (consultato in data 16/09/2022).

Corte di Giustizia dell'Unione Europea (CURIA) - Sentenza della Corte (Quarta Sezione) del 13 febbraio 2014, https://curia.europa.eu/juris/document/document.jsf;jsessionid=9841C0B072CCE5035475DA3AB7902FD0?text=&docid=147847&pageIndex=0&doclang=it&mode=lst&dir=&occ=first&part=1&cid=9224624 (consultata in data 16/09/2022).

Sitografia

European Commission - ERIC, https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/european-research-infrastructures/eric_en (visitato in data 22/05/2022).

Ministero dell'Istruzione, dell'Università e della Ricerca - ESFRI, https://www.istruzione.it/archivio/web/ricerca/ricerca-internazionale/esfri.html (visitato in data 23/05/2022).

APRE - Horizon Europe - Home, https://horizoneurope.apre.it (visitato in data 29/05/2022).

CLARIN ERIC - Home, https://www.clarin.eu (visitato in data 04/06/2022).

CLARIN ERIC - CLARIN Technology: An Introduction, https://www.clarin.eu/content/clarin-technology-introduction (visitato in data 04/06/2022).

CLARIN ERIC - CLARIN Centres, https://www.clarin.eu/content/clarin-centres (visitato in data 04/06/2022).

CLARIN ERIC - The Knowledge Infrastructure, https://www.clarin.eu/content/knowledge-infrastructure (visitato in data 04/06/2022).

CLARIN ERIC - Easy-to-Use Language Resources, https://www.clarin.eu/content/language-resources (visitato in data 04/06/2022).

CLARIN ERIC - Resource Families, https://www.clarin.eu/resource-families (visitato in data 04/06/2022).

CLARIN ERIC - Data, https://www.clarin.eu/content/data (visitato in data 04/06/2022).

CLARIN ERIC - About CLARIN in the Preparatory Phase, https://www.clarin.eu/content/about-clarin-preparatory-phase (visitato in data 11/06/2022).

CLARIN ERIC - CLARIN ERIC Statutes, https://www.clarin.eu/content/clarin-eric-statutes (visitato in data 11/06/2022).

CLARIN ERIC - Governance, https://www.clarin.eu/content/governance (visitato in data 11/06/2022).

CLARIN ERIC - Vision and Strategy, https://www.clarin.eu/content/vision-and-strategy (visitato in data 11/06/2022).

CLARIN ERIC - Participating Consortia, https://www.clarin.eu/content/participating-consortia (visitato in data 11/06/2022).

CLARIN ERIC - Certified B-centres, https://www.clarin.eu/content/certified-b-centres (visitato in data 11/06/2022).

CLARIN ERIC - Overview CLARIN Centres, https://www.clarin.eu/content/overview-clarin-centres (visitato in data 11/06/2022).

CLARIN ERIC - CLARIN Centre Registry, https://centres.clarin.eu (visitato in data 11/06/2022).

CLARIN ERIC - Overview of CLARIN K-centres, ordered by acronym, https://vonweber.nl/cgi/kcentres_page.cgi (visitato in data 11/06/2022).

CLARIN ERIC - Virtual Language Observatory (VLO), https://www.clarin.eu/content/virtual-language-observatory-vlo (visitato in data 25/06/2022).

CLARIN ERIC - Federated Content Search (CLARIN-FCS) - Technical Details, https://www.clarin.eu/content/federated-content-search-clarin-fcs-technical-details (visitato in data 25/06/2022).

CLARIN ERIC - Language Resource Switchboard, https://www.clarin.eu/content/language-resource-switchboard (visitato in data 25/06/2022).

CLARIN ERIC - CLARIN in EU Projects, https://www.clarin.eu/content/clarin-eu-projects (visitato in data 25/06/2022).

CLARIN ERIC - CLARIN in past EU Projects, https://www.clarin.eu/content/clarin-past-eu-projects (visitato in data 25/06/2022).

CLARIN ERIC - CLARIN Funding Hub, https://www.clarin.eu/funding (visitato in data 25/06/2022).

CLARIN ERIC - CLARIN Annual Conference, https://www.clarin.eu/content/clarin-annual-conference (visitato in data 25/06/2022).

CLARIN ERIC - Tour de CLARIN, https://www.clarin.eu/Tour-de-CLARIN (visitato in data 25/06/2022).

Zotero - CLARIN Library, https://www.zotero.org/groups/562080/clarin/library (visitato in data 07/05/2022).

CLARIN-IT - Home, https://www.clarin-it.it (visitato in data 25/06/2022).

CLARIN-IT - Virtual Language Observatory, https://www.clarin-it.it/en/content/virtual-language-observatory (visitato in data 25/06/2022).

CLARIN-IT - Federated Content Search, https://www.clarin-it.it/en/content/federated-content-search (visitato in data 25/06/2022).

CLARIN-IT - Participation in Innovative and Strategic Projects, https://www.clarin-it.it/en/content/participation-innovative-and-strategic-projects (visitato in data 25/06/2022).

CLARIN-IT - Support to Projects and Events of the Sector of Social Sciences and Humanities, https://www.clarin-it.it/en/content/support-projects-and-events-sector-social-sciences-and-humanities (visitato in data 25/06/2022).

CLARIN-IT - Cretan Institutional Inscriptions (Web site), https://www.clarin-it.it/en/content/cretan-inscriptions (visitato in data 25/06/2022).

ILC4CLARIN - Cretan Institutional Inscriptions (Web application), https://ilc4clarin.ilc.cnr.it/cretaninscriptions/en (visitato in data 25/06/2022).

ILC4CLARIN Repository - Cretan Institutional Inscriptions Dataset, https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-548 (visitato in data 25/06/2022).

ILC4CLARIN Repository - Cretan Institutional Inscriptions (Web application), https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/handle/20.500.11752/OPEN-550 (visitato in data 25/06/2022).

DiPText-KC - Home, https://diptext-kc.clarin-it.it (visitato in data 29/01/2022).

DiPText-KC - About, https://diptext-kc.clarin-it.it/about (visitato in data 29/01/2022).

DiPText-KC - Bibliographic Resources, https://diptext-kc.clarin-it.it/knowledge/bibliographic-resources (visitato in data 29/01/2022).

Zotero - DiPText-KC Library, https://www.zotero.org/groups/4521720/diptext/library (realizzata dal candidato nell'ambito di un tirocinio curriculare svolto presso il CNR-ILC e pubblicata in data 22/04/2022).

DARIAH-EU - Home, https://www.dariah.eu (visitato in data 17/06/2022).

DARIAH-EU - DARIAH in a Nutshell, https://www.dariah.eu/about/dariah-in-nutshell (visitato in data 17/06/2022).

DARIAH-EU - History of DARIAH, https://www.dariah.eu/about/history-of-dariah (visitato in data 17/06/2022).

DARIAH-EU - Mission & Vision, https://www.dariah.eu/about/mission-vision (visitato in data 19/06/2022).

DARIAH-EU - Members and Partners, https://www.dariah.eu/network/members-and-partners (visitato in data 19/06/2022).

DARIAH-EU - Organisation and Governance, https://www.dariah.eu/about/organisation-and-governance (visitato in data 19/06/2022).

DARIAH-EU - Open Science, https://www.dariah.eu/activities/open-science (visitato in data 19/06/2022).

DARIAH-EU - Projects, https://www.dariah.eu/activities/projects-list (visitato in data 19/06/2022).

DARIAH-EU - EOSC Future, https://www.dariah.eu/activities/projects-and-affiliations/eosc-future (visitato in data 19/06/2022).

DARIAH-EU - CLS INFRA, https://www.dariah.eu/activities/projects-and-affiliations/cls-infra (visitato in data 21/06/2022).

DARIAH-EU - SSHOC, https://www.dariah.eu/activities/projects-and-affiliations/sshoc (visitato in data 21/06/2022).

DARIAH-EU - DiXiT, https://www.dariah.eu/activities/projects-and-affiliations/dixit (visitato in data 21/06/2022).

DARIAH-EU - Hypotheses.org – Academic Blogs, https://www.dariah.eu/tools-services/tools-and-services/tools/hypotheses-org-academic-blogs (visitato in data 21/06/2022).

DARIAH-EU - Bibliography "Doing Digital Humanities", https://www.dariah.eu/tools-services/tools-and-services/tools/bibliography-doing-digital-humanities (visitato in data 23/06/2022).

DARIAH-EU - DARIAH-DE Working Papers, https://www.dariah.eu/tools-services/tools-and-services/tools/dariah-de-working-papers (visitato in data 23/06/2022).

DARIAH-EU - Training and Education, https://www.dariah.eu/activities/training-and-education (visitato in data 23/06/2022).

Università di Pisa - AIUCD 2021, https://aiucd2021.labcd.unipi.it (visitato in data 25/06/2022).

CNR-ILC CoPhiLab - https://cophilab.ilc.cnr.it/euporia2021 (visitato in data 25/06/2022).

DiXiT - About, https://dixit.uni-koeln.de/about (visitato in data 08/07/2022).

Università di Pisa - Lo European Open Science Cloud (EOSC), https://www.unipi.it/index.php/open-science/item/18013-24-lo-european-open-science-cloud-eos (visitato in data 19/07/2022).

GitHub - Digital Humanities Taxonomy Group - TaDiRAH - Taxonomy of Digital Research Activities in the Humanities, https://github.com/dhtaxonomy/TaDiRAH/blob/master/introduction.md (visitato in data 27/06/2022).

TaDiRAH - Taxonomy of Digital Research Activities in the Humanities - Home, https://tadirah.info (visitato in data 27/06/2022).

Aziona - API cosa sono e come funzionano, https://www.azionadigitale.com/api-cosa-sono-e-come-funzionano (visitato in data 10/08/2022).

Openapi - Cos'è una chiamata API, https://openapi.it/blog/cosa-e-chiamata-api.html (visitato in data 10/08/2022).

MRW.it - Il Web Scraping in PHP, https://www.mrw.it/php/web-scraping-php_7568.html (visitato in data 12/06/2022).

Sviluppo PHP - Come implementare un Web scraper in PHP?, https://php.yocker.com/come-implementare-un-web-scraper-in-php.html (visitato in data 12/06/2022).

QUISH - 8 fantastiche librerie e strumenti di Web Scraping PHP, https://it.quish.tv/8-awesome-php-web-scraping-libraries (visitato in data 12/06/2022).

Stack Overflow - How to implement a Web scraper in PHP?, https://stackoverflow.com/questions/26947/how-to-implement-a-web-scraper-in-php (visitato in data 07/07/2022).

PHP.net - PHP Manual, https://www.php.net/manual/en/index.php (visitato in data 09/07/2022).

Creative Commons Italia - Le licenze, https://creativecommons.it/chapterIT/index.php/license-your-work (visitato in data 17/09/2022).

Ringraziamenti

Un ringraziamento particolare va al Dottor Angelo Mario Del Grosso e alla Dottoressa Paola Baroni, per avermi guidato nello svolgimento del tirocinio curricolare presso il CNR-ILC e per avermi sostenuto nel processo di compimento del percorso accademico.

Una menzione d'onore spetta al Dottor Federico Boschetti, Responsabile dell'Unità di Ricerca del CNR-ILC presso il VeDPH di Venezia e Referente del CLARIN K-Centre DiPText-KC, per i suoi preziosi consigli sia nel corso del tirocinio curricolare sia nella produzione del presente elaborato.

Un ringraziamento speciale va alla mia famiglia. Nessuna parola sarà mai abbastanza per esprimere la mia gratitudine a mia madre e a mio padre, per il loro costante sostegno e per tutti i sacrifici che hanno fatto per me: senza di loro, tutto questo non sarebbe stato possibile. Grazie a nonna Carla, per essere semplicemente fantastica. Grazie ai miei zii e a Daniele, che è stato per me come un fratello maggiore e a cui auguro una vita rosea e felice a fianco di Camilla, a cui estendo il mio ringraziamento.

Un grazie di cuore agli amici di una vita, a quelli che mi conoscono meglio di come mi conosca io, a chi è partito, a chi partirà e a chi è sempre presente: auguro a tutti loro di avere sempre il vento in poppa e il coraggio di non arrendersi mai.

Appendice A - La funzione *scrape*

```
function scrapeIssuesYear($archiveURL){
   $queryString = http_build_query([
      'access_key' => '7387b2bf6eccddc112606fb96e37cb38',
      'url' => $archiveURL,
                  ]);
      // API URL con query string
   $apiURL = sprintf('%s?%s', 'http://api.scrapestack.com/scrape',
$queryString);
   $dshCompletamento = "https://academic.oup.com";

   $ch = curl_init();
   curl_setopt($ch, CURLOPT_URL, $apiURL);
   curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
   $html = curl_exec($ch);
   curl_close($ch);

   $dom = new DOMDocument();
   @ $dom->loadHTML($html);
   $xpath = new DomXPath($dom);

   $resultsVolume = [];
   $arrAllIssuesLink = [];


   //YEAR ISSUE
   $results = [];
  $results['year'] = $xpath->query('//div[@class="widget
widget-IssueYears widget-instance-OUP_Issues_Year_List"]//div//a');
  $results['link'] = $xpath->query('//div[@class="widget
widget-IssueYears
widget-instance-OUP_Issues_Year_List"]//div//a//@href');
    $arrYear = [];
    for($x=0; $x < $results['link']->length;$x++){
        $year = $results['year']->item($x)->textContent;
        $link = $results['link']->item($x)->textContent;
           if($year > '2015'){
              $arrYear[] = $dshCompletamento.$link;
           }
    }
    return $arrYear;
}
```

Appendice B - Sistemazione e correzione dei dati

```
$final_year = '';

if( strpos($year,'Selected Papers')){
    $year = explode("|",$year);
    $newstr = explode("Selected Papers",$year[1]);
    $final_year = preg_replace('/[^0-9]/', '', $newstr[1]);
    if(strlen($final_year) == 8){
        $splitstring1 = substr($final_year, 0, floor(strlen($final_year)
/ 2));
        $splitstring2 = substr($final_year, floor(strlen($final_year) /
2));
        if (substr($splitstring1, 0, -1) != ' ' AND
substr($splitstring2, 0, 1) != ' '){
          $middle = strlen($splitstring1) + strpos($splitstring2, ' ');
        } else {
          $middle = strrpos(substr($final_year, 0,
floor(strlen($final_year) / 2)), ' ');
        }
        $string1 = substr($final_year, 0, $middle);
        $string2 = substr($final_year, $middle);
        $final_year = $string2;
 }
}
 else if(strpos($year,'Reaching') ){
            $year = explode("|",$year);
            $newstr = explode("Reaching",$year[1]);
            $final_year = preg_replace('/[^0-9]/', '', $newstr[0]);
        }
else{
            $year = explode("|",$year);
            $final_year = preg_replace('/[^0-9]/', '', $year[1]);

        }
 if(strpos($title,"          ")){
    $title = str_replace("      ","",$title);
    if(strpos($title,"   ")){
       $title = str_replace("   "," ",$title);
 }
}
```

Appendice C - La funzione *add*

```
function add_all_Issues(){
  global $conn;
  $all_issues_archive = 'https://academic.oup.com/dsh/issue-archive';
  $All_issues2015to2020 = DOSCRAPEOFALLISSUE($all_issues_archive);

   $arr = [];
   $i=0;
   foreach($All_issues2015to2020 as $k =>$val_arr){
      foreach($val_arr as $key =>$val){
            $arr[$i]['title']  = $val['title'];
            $arr[$i]['link']   = $val['link'];
            $arr[$i]['volume']       = $val['volume'];
            $arr[$i]['year']   = $val['year'];
            $arr[$i]['authors']      = $val['authors'];
            $i++;
      }
    }
   foreach($arr as $key =>$val){
      $title = $val['title'];
      $link = $val['link'];
      $vol  = $val['volume'];
      $year = $val['year'];
      $author = $val['authors'];

       $sql_check = "SELECT `art_title` FROM articles WHERE `art_title`=
'$title' AND  art_table_id = 2 ";
       $result = mysqli_query($conn, $sql_check);
       $art = mysqli_fetch_all($result, MYSQLI_ASSOC);

    if(empty($art)){
      $query = "INSERT INTO articles(art_id, art_title, art_author,
      art_link, art_year, art_topic, art_volume, art_type, art_table_id
      VALUES('','$title', '$author' ,'$link','$year',null,'$vol','IS',2
      )";
      mysqli_query($conn, $query);
   }
 }
 $day = date("Y-m-d");
 $query_logs =   "INSERT INTO logs(`id`, `last_update`, `table_id`,
`art_type`) VALUES('','$day', 2,'IS')";
 mysqli_query($conn, $query_logs);
}
```

Appendice D - Attivazione della funzione *add* tramite *click*

```
$("#addAdvArt").click(function(){
          var hiddenDSH = $('#tabdsh').val();
          $.ajax({
            url: './rpc/manage_articles.php',
            type: 'post',
            data: {
              tab:hiddenDSH,
              type:'AD'
            },
            beforeSend: function(){
              $('#spinner').show();
              $('#container-spinner').show();
              $('#mainDIV').css('opacity', 0.1);
            },
            success: function(response){
              openModal();
            },
            complete:function(data){
              $('#spinner').hide();
              $('#container-spinner').hide();
              $('#mainDIV').css('opacity', 1);
            }
          });
```

Appendice E - Creazione della tabella nel sito Web

```
<table  id="myDatatable"  class="table”  table-striped"  cellspacing="0"
width="100%">
       <thead>
        <tr>
          <th style="display:none;">#</th>
          <th>Author</th>
          <th>Title</th>
          <th>Year</th>
          <th>Volume</th>
        </tr>
      </thead>

      <tfoot>
        <tr>
         <th style="display:none;">#</th>
         <th>Author</th>
         <th>Title</th>
         <th>Year</th>
         <th>Volume</th>
       </tr>
      </tfoot>

     <tbody>
         <?php foreach ( $result as $art_id=>$value ) {?>
            <tr>
              <td style="display:none;" class="hidden-xs">
                  <?= $value['art_id'] ?>
              </td>
              <td>
                   <?= $value['art_author']?>
              </td>
              <td  class="focus">
                 <a href="<?= $value['art_link'] ?>">
                    <?= $value['art_title'] ?>
                 </a>
              </td>
              <td>
                  <?= $value['art_year']  ?>
              </td>
              <td>
                  <?= $value['art_volume'] ?>
              </td>
            </tr>
         <?php }?>
     </tbody>
</table>
```

Appendice F - Filtro avanzato

```
function filtroAvanzatoGS($post){
      global $conn;
      if(!empty($post['autore'])){
            $autore = $post['autore'];
            $AND_aut = "AND art_author LIKE '%$autore%' ";
      }else{
            $AND_aut = '';
            }
      if(!empty($post['titolo'])){
            $titolo = $post['titolo'];
            $AND_tit = "AND art_title LIKE '%$titolo%' ";
      }else{
            $AND_tit = '';
      }
  if(!empty($post['annoDA']))
       {$annoDA = $post['annoDA'];}else{$annoDA = 0;}
  if(!empty($post['annoA']))
       {$annoA = $post['annoA'];}else{$annoA = 0;}

      if(!empty($post['anno'])){
            $AND_year = 'AND (';
            $count = count($post['anno'])-1;
            foreach($post['anno'] as $k => $anno){
                  if($k < $count){
                        $AND_year .= 'art_year = '.$anno.' OR ';
                  }else if($k == $count){
                        $AND_year .= 'art_year = '.$anno.' ';
            }}
            $AND_year .= ')';
      }else{
            $AND_year = '';
            }
      if( $annoDA && $annoA){
         if( $annoA < $annoDA){
            $AND_DA_A = "AND art_year BETWEEN $annoA AND $annoDA ";
         }else{
            $AND_DA_A = "AND art_year BETWEEN $annoDA AND $annoA ";
         }
      }else{
            $AND_DA_A = '';
      }
      $sql = "SELECT *
            FROM articles
            WHERE art_table_id = 4
            $AND_aut
            $AND_tit
            $AND_year
            $AND_DA_A";

      $result = mysqli_query($conn, $sql);
      $articles = mysqli_fetch_all($result, MYSQLI_ASSOC);
      return $articles;}
```