

SERICA

Una banca dati per le relazioni Occidente-Oriente

Andrea Balbo

Università di Torino

andrea.balbo@unito.it

Abstract

Il contributo presenta il sito *SERICA* e le sue opportunità di utilizzo e di sviluppo per le ricerche linguistiche e per le indagini testuali relative a un *corpus* di testi in progressivo incremento e contenente prevalentemente scritti latino, ma anche qualche caso di testi spagnoli e russi.

Parole chiave: Cina; Gesuiti; Latino; Linguistica; Linguistica latina; Banca dati.

The contribution presents the SERICA website and its opportunities for use and development in linguistic research and textual investigations relating to a corpus of texts that is progressively expanding and consists predominantly of Latin writings, with some instances of Spanish and Russian texts as well.

Keywords: China; Jesuits; Latin; Linguistics; Latin Linguistics; Database.

1. Il progetto *SERICA*

SERICA è l'acronimo di *Sino-European Religious Intersections in Central Asia. Interactive Texts and Intelligent Networks*, progetto che ha ottenuto un finanziamento dal Ministero dell'Università italiano nell'ambito del bando FISR (Fondo integrativo speciale per la ricerca) 2019 per la durata di due anni, dal 2021 al 2023. *SERICA* ha previsto l'attività di due gruppi di ricerca, il primo, basato nell'Università di Pisa e guidato dal P. I. Chiara Ombretta Tommasi (Storia delle religioni), il secondo, incardinato nell'Università di Torino e guidato dal P. I. Andrea Balbo (Lingua e letteratura latina). Essi hanno coinvolto ben 20 ricercatori e studiosi, 7 assegnisti di ricerca post-dottorato e numerosi collaboratori temporanei e hanno lavorato in modo molto stretto sia con il Dipartimento di Informatica dell'Università di Pisa sia con il gruppo di ricerca di Francesco Stella a Siena, che gestisce il progetto *ELA (Eurasian Latin Archive)*.¹ Il progetto ha contribuito a portare in modo massiccio all'attenzione della comunità scientifica gli studi eurasiatici o gli studi sulle Vie della Seta, oggi annoverati tra i temi più vivaci e stimolanti della storia globale. Tali studi hanno spesso messo in luce il ruolo dei mercanti itineranti e dei missionari che, a partire dalla

¹ <https://ela.unisi.it/>.

Tarda Antichità, iniziarono a percorrere i primi itinerari culturali che unirono l'Asia orientale e l'Europa in uno scambio strutturale e permanente, che raggiunse il suo apice tra il XVI e il XVIII secolo, con la produzione di un enorme numero di testi e la creazione di biblioteche multilingui, in particolare in Cina. Queste opere miravano a diffondere in Estremo Oriente elementi dell'educazione classica europea e, al contempo, a trasmettere in Europa le prime informazioni di ampia diffusione sulla geografia, sui costumi e sui fondamenti culturali delle civiltà dell'Asia orientale. All'estremità occidentale delle Vie della Seta, in Anatolia e nel Levante, le complesse interazioni tra il cristianesimo occidentale e orientale e il mondo musulmano nel Medioevo hanno anch'esse ricevuto una notevole, seppur disomogenea, attenzione da parte degli studiosi. Una rivalutazione delle fonti che testimoniano le interazioni tra cristiani e musulmani (ad esempio le opere di carattere etnografico) è apparsa quanto mai necessaria, al fine di chiarirne il pubblico di riferimento, la più ampia circolazione e la ricezione successiva. Questo processo multiforme di comunicazione e di scambio culturale è stato oggetto di numerosi studi storici, quasi sempre basati su una selezione limitata di documenti e raramente affrontati dal punto di vista dell'intersezione linguistica tra le civiltà coinvolte, né dalla prospettiva dei contatti orientali con l'Europa e l'Occidente. L'obiettivo del progetto è stato precisamente quello di riunire un'équipe interdisciplinare di esperti e, al tempo stesso, di promuovere ulteriori collaborazioni nazionali e internazionali, sottolineando le connessioni con altri progetti affini. Attraverso la selezione di alcuni tra i più significativi casi di studio in un arco cronologico molto ampio (400 a.C.-1700 d.C.), abbiamo analizzato le modalità con cui libri e manoscritti giunsero in Europa e talvolta furono ulteriormente diffusi tramite traduzioni. La circolazione di questi testi testimonia non solo relazioni diplomatico-commerciali, ma anche un'interazione osmotica tra diverse istanze, fedi, sistemi sociali e universi culturali che si muovevano lungo tali itinerari. Considerando le cose materiali come punti di accesso alla storia, la piattaforma digitale del progetto e la mappatura virtuale dei percorsi storico-geografici menzionati nei libri o effettivamente seguiti dalla loro circolazione intendono mettere in evidenza, nel senso più ampio, il contatto ininterrotto tra Asia ed Europa. La scelta di un lungo periodo storico e di una vasta area geografica consente l'adozione di un programma transdisciplinare, che favorisce una revisione delle metodologie tradizionali e mira a una lettura non eurocentrica del passato: entrambi questi aspetti permettono di riflettere su e rispondere a questioni cruciali per la società contemporanea. Di conseguenza, la novità dell'approccio di *SERICA* può essere riassunta in un duplice aspetto: da un lato, esso presenta e rende disponibili nuovi documenti, spesso inediti o poco noti, la maggior parte dei quali non è mai stata tradotta in italiano o in altre lingue moderne; dall'altro, ciò è stato reso possibile anche grazie al ruolo centrale delle ICT, che consentono la valutazione e l'interpretazione dei dati combinandoli in un *corpus* unitario, trasversale e dinamico, capace di sfruttare appieno le potenzialità delle tecnologie informatiche. Si sono così aperte nuove prospettive diacroniche negli studi comparativi. Uno degli obiettivi del progetto *SERICA* è lo sviluppo di una biblioteca digitale,² che mira a includere testi relativi alle rotte dell'Asia centrale tra la Cina e l'Europa. Molti di questi testi furono scritti in latino da missionari gesuiti che viaggiarono in Cina tra il XVI e il XVIII secolo e redassero numerose opere per diffondere in Europa la storia e la cultura cinesi, insieme al pensiero confuciano. Il sito presenta un'interfaccia utente facilmente usabile ed efficace.

Mi occuperò qui di seguito della parte testuale, viste le finalità del convegno *Digital Latin*. Il caricamento dei documenti sul sito web di *SERICA* consente agli utenti di effettuare ricerche lessicali mirate, accedendo a testi che in precedenza erano stati digitalizzati e diffusi esclusivamente in formati non interrogabili. Questo processo si articola in diverse fasi. In primo

² Sito web: <https://SERICA.unipi.it/>.

luogo, il documento – normalmente libri a stampa in pdf disponibili *online* o scansionati³ – a seconda del suo aspetto o delle modalità di digitalizzazione, viene trascritto o elaborato mediante un software OCR. Successivamente, il testo è sottoposto a correzione linguistica e tipografica e a normalizzazione, al fine di eliminare eventuali errori. In questa fase, per quanto riguarda i testi latini, le abbreviazioni vengono sciolte nei corrispondenti termini completi, il grafema <j> viene trascritto come <i> e tutti gli accenti sulle vocali vengono rimossi (ad es. *Proëmialis Declaratio*, p. IX: *eò etiam magis, quòd placuit > eo etiam magis, quod placuit*), così da facilitare una ricerca lessicale esatta nel database. Completata questa fase, il testo viene convertito in formato XML, marcato secondo lo standard TEI e caricato sul sito. Ad oggi, il *corpus* include 53 documenti interrogabili, per un totale di 472.194 parole e 33.852 lemmi. La maggioranza dei testi è in latino, con la presenza di qualche testo in spagnolo e in russo, ma il sito è predisposto anche per l'accoglienza di testi in cinese. Le opere riguardano principalmente la Cina e il Giappone, ma anche l'India. I testi sono presentati come in Fig. 1 ed è possibile anche individuare brevi informazioni sul nome dell'autore e nella sezione Informazioni linguistiche il numero delle parole e dei lemmi.



Fig. 1 – Pagina contenente un testo di SERICA

Il *corpus* di testi presente attualmente è elencato nella Tab. 1.

Ad virum nobilem de cultu Confucii philosophi et progenitorum apud Sinas
Algoritmi de numero Indorum
Anotaciones sobre las misiones, Roma, c. agosto de 1558
Brevis Relatio
Brevis Relatio de Numero, Et Qualitate Christianorum apud Sinas

³ Non abbiamo ancora preso in esame la possibilità di utilizzare manoscritti, ma il recente convegno del PRIN MARE a Enna (18-19 dicembre 2025) ha aperto ulteriori prospettive su questo tema.

Brevis relatio eorum, quæ spectant ad declarationem Sinarum imperatoris Kamhi circa cæli, Cumfucii, et avorum cultum

Catechismus cristianae fidei

Confucius Sinarum Philosophus sive Scientia Sinensis Latine exposita – Epistula praeceptorum

Confucius Sinarum Philosophus sive Scientia Sinensis Latine exposita – Philosophorum sinensium Principis Confucii vita

Confucius Sinarum Philosophus sive Scientia Sinensis Latine exposita – Proemialis Declaratio

Confucius Sinarum Philosophus sive Scientia Sinensis Latine exposita – Scientiae Sinicae Liber Primus

Consultatio Macaensis anno 1645 habita de forma baptismi annamitici

De antiquitate Sinarum

De horis sinicis

De magno Sinarum imperio

Dissertationes duae de Mose Mardeno, una

ELOGIUM REVERENDI PATRIS FERDINANDI VERBIEST SOCIETATIS IESU

Epistola ad Cardinales S.C.P.F., 5 nov. 1672

Epistola ad Episcopos Tunkinenses, 29 Ian. 1675

Epistola ad Episcopos Tunkinenses, 5 febr. 1674

Epistola ad Illm. Bernardinum Della Chiesa, 15 Dec. 1684

Epistola de mensuris et ponderibus serum seu sinensium

Epistola Fratris Andreas de Perusio de Ordine Minorum fratrum reverendo patri fratri Guardiano Perusini Conventus

Epistola II

Epistola III

EPISTOLA ILLUSTRISSIMI DOMINI GREGORII LOPEZ AD R. P. PHILIPPUM COUPLÉT Societatis IESU Procuratorem Sinarum tunc Romae degentem

Epistolium ex regno Sinarum ad mathematicos Europaeos missum cum commentario Johannis Kepler

Epistolium Terrentii cum commentariuncula Kepleri

Flora Sinensis

Histoire générale des Huns, des Turcs, des Mogols, et des autres Tartares occidentaux

Historiarum Indicarum Libri XVI

Historiarum Indicarum Libri XVI (tit.)

Historica notitia rituum ac ceremoniarum sinicarum

<p>Icon regia Sinarum nunc regnantis ex gallico versa</p> <p>Ignatius De Loyola Francisco Xaverio et superioribus S.I. in Oriente, Roma 7 Iulii 1550</p> <p>Ignatius De Loyola P. Francisco Xaverio, in Indiam, Roma 10 Octobris 1549</p> <p>Informatio pro Provincia Japponica [1646]</p> <p>КНИГА, А ВЪ НЕЙ ПИСАНО ПУТЬШЕСТВИЕ ЦАРСТВА СИБИРСКАГО ОТЪГОРОДА ТОБОЛСКА И ДО САМОГО РУБЕЖА ГОСУДАРСТВА КИТАЙСКОГО, ЛѢта 7183 мѣсяца мая въ 3-й день</p> <p>Melchor Nunes Barreto a Diego Laínez, Cochín, 13 de enero de 1558</p> <p>Novissima sinica</p> <p>[P. Antonius Criminalis S.I.] P. Ignatio de Loyola, Romam, e promontorio Comorin 4 decembris 1548, secunda via</p> <p>P. Ignatius De Loyola P. Paulo S.I, Goam, Roma 20 Novembris 1547</p> <p>P. Ioannes de Polanco S.I. ex Comm. P. Nicolao Lancillotto S.I, Goam, Roma 22 Novembris 1547 (II)</p> <p>P. Ioannes de Polanco S.I. ex Comm. [P. Nicolao Lancillotto S.I, Goam], Roma 22 Novembris 1547 (III)</p> <p>P. Ioannis De Polanco S.I. Excerpta e litteris indicis de electione patris quadros in provincialem [Goa Ianuario 1556] SCRIPTIS</p> <p>P.P. Ioannes de Polanco S.I. ex Comm. P. Nicolao Lancillotto S.I, Goam, Roma 22 Novembris 1547</p> <p>Relatio Fr. Benedicti Poloni</p> <p>Relatio insignis miraculi a S. Francisco Xaverio societatis Iesu</p> <p>Relatio sepulturae magno Orientis apostolo S. Francisco Xaverio erectae in insula Sanciano anno saeculari MDCC</p> <p>Relazione della China cavata da un ragionamento tenuto col padre Giovanni Grueber della compagnia di Gesù nel suo passaggio per Firenze l'anno 1665</p> <p>Responsa ad D. Carolum Maigrot, 15 et 22 Mart. 1689</p> <p>Responsum ad intimationem D. Maigrot 30 nov. 1689</p> <p>Responsum ad Procuratorem Generalem, 11 Dec. 1689</p>

Tab. 1 – Elenco dei testi attualmente presenti in *SERICA*

Dal punto di vista linguistico, va ricordato come siano state implementate recentemente alcune funzioni che, in attesa di quelle progettate precedentemente, cominciano a offrire possibilità di analisi linguistiche. Sono infatti disponibili, oltre alle funzioni di lettura e di scaricamento, le seguenti attività di ricerca:

- a. l'opzione di filtraggio per Autori

- b. la ricerca di parole all'interno del *corpus* (o dei singoli autori) con operatori *booleani* che consentono anche l'inserzione di parti di parola; nella maschera si può indicare anche il numero di parole interposte, fatto che consente la ricerca di sintagmi; in via di implementazione sono le funzioni “Eccezioni” e “Ricerca Sperimentale”;
- c. la funzione “Analizza il *corpus*” e “Analizza il documento”, che consente di identificare nel *corpus* o in un singolo documento i lemmi e i *token* con e senza *stop words*. Nell'analisi del *corpus* risulta possibile identificare il numero di parole di ogni testo del *corpus* e la loro percentuale all'interno del *corpus* totale portandosi con un *click* su ogni singolo documento, è possibile evidenziare un ulteriore livello di dati:
 - i. il lemma
 - ii. la *term frequency* all'interno del documento
 - iii. la *term frequency* nei documenti
 - iv. il rapporto tf/idf^4

Di particolare interesse è la funzione “Compara”, che si realizza attraverso la costruzione di gruppi di testi selezionati (vd. Fig. 2).

⁴ TF-IDF (*Term Frequency-Inverse Document Frequency*) è una misura numerica utilizzata nel reperimento delle informazioni e nell'analisi testuale per valutare quanto una parola sia importante all'interno di un documento rispetto a una collezione di documenti (un *corpus*). La frequenza del Termine (*TF*) misura quante volte un termine compare in un determinato documento. L'idea di base è che i termini che compaiono più frequentemente in un documento siano più rilevanti per il suo contenuto. La frequenza inversa del documento (*IDF*) misura quanto un termine sia raro nell'intero *corpus*. Le parole che compaiono in molti documenti (come le parole funzione) ricevono un peso minore, mentre quelle che compaiono in pochi documenti ricevono un peso maggiore. Quando il valore di *TF-IDF* è alto, indica che un termine compare frequentemente in un documento specifico (alta frequenza del termine), ma compare raramente nel resto del *corpus* (alta frequenza inversa del documento). Di conseguenza, il termine è considerato altamente informativo e caratteristico di quel documento specifico, poiché contribuisce a distinguerlo dagli altri testi della collezione. Un punteggio *TF-IDF* elevato segnala quindi che il termine ha un forte potere discriminante ed è particolarmente utile per attività quali il reperimento delle informazioni, l'estrazione di parole chiave, il ranking dei documenti e la classificazione testuale. Quando il valore di *TF-IDF* è basso, indica che un termine compare raramente nel documento (bassa frequenza del termine), e/o compare frequentemente nel *corpus* (bassa frequenza inversa del documento). In pratica, ciò significa che il termine ha scarso valore discriminante o informativo per quel documento specifico. Tali termini sono spesso parole comuni o generiche che non aiutano a distinguere un documento dagli altri, oppure parole solo marginalmente rilevanti per il contenuto del testo: i termini non sono un indicatore forte dell'argomento del documento e sono generalmente meno utili per compiti come la classificazione dei documenti, il recupero delle informazioni o l'estrazione di parole chiave.

Fig. 2 – Prima fase della creazione di due *corpora* da comparare

Ipotizziamo una costruzione come quella in Fig. 3, realizzata spuntando i titoli delle opere.

Fig. 3 – Seconda fase della creazione di due *corpora* da comparare

Cliccando sulla funzione “Compara” i gruppi, si ottengono visualizzazioni comparate come in Fig. 4:

lemma	term frequency nel documento	in documenti	tf/idf
-ne	2280	47	182.5
a-noiso	6	5	12.56
ab	4351	48	258.55
abusivos	8	4	18.73
ablico	6	6	12.03

lemma	term frequency nel documento	in documenti	tf/idf
-ne	2280	47	182.5
ab	4351	48	258.55
abreuiacio	1	1	3.26
abreuiatio	2	2	5.71
absoque	55	20	49.87

Fig. 4 – Fase conclusiva e comparazione di due *corpora*

Da qui si può ripartire con una serie di altre analisi che utilizzano i dati ricavati anche a livello sintattico⁵.

2. Le prospettive future

Il lavoro sul sito non è concluso: a parte l'implementazione, esso sarà strutturato secondo diverse fasi di progettazione prototipale (dal *wireframing* al *graphic design*). Tali fasi includono la definizione delle categorie generali di utenti (*proto-persona*), la costruzione dei *use case* e delle *user stories*, la descrizione dei percorsi di navigazione (*user journey*) e l'analisi dell'efficacia della proposta tramite interviste e test sugli utenti. Una volta completata, la biblioteca digitale sarà dotata di strumenti innovativi di *semantic web* e di intelligenza artificiale, nonché una serie di mappe interattive diacroniche che utilizzano strumenti di geolocalizzazione. I dati sono stati raccolti in un database interattivo, con una mappatura visiva dei testi e degli oggetti che documentano gli scambi reciproci, principalmente dall'Oriente verso l'Occidente. Il risultato finale sarà una mappa diacronica interattiva, navigabile e “parlante” delle diverse rotte, nella quale immagini e dettagli storici contribuiranno a ricostruire nel modo più accurato possibile il contesto delle conoscenze che un viaggiatore avrebbe potuto possedere. Questa biblioteca digitale è in fase di realizzazione mediante *suites* che includono funzioni quali il caricamento di testi e metadati, l'arricchimento semantico e la gestione dei dati. Tale base di conoscenza rappresenterà un modello concreto (e il primo in Italia) per la visualizzazione interattiva dei dati relativi alle relazioni storiche tra i diversi Paesi orientali e l'Europa.⁶ Inoltre, il contenuto librario e testuale è divenuto parte integrante del sistema. Un'attenzione particolare è rivolta ai processi di analisi linguistica (attraverso tecniche di *Natural Language Processing*), al fine di normalizzare i testi ed estrarre la maggiore quantità possibile di informazioni strutturate da utilizzare come input per interrogazioni avanzate e analisi di intelligenza artificiale. Un ambito di particolare interesse è la diversità temporale dell'origine delle fonti, che rende l'analisi NLP particolarmente complessa, poiché gli strumenti disponibili prendono generalmente in considerazione solo le versioni moderne delle lingue e non quelle antiche o classiche. In particolare, la natura ibrida del progetto (biblioteca digitale-mappa interattiva-database) lo rende un modello pilota nel panorama della ricerca italiana sugli studi delle lingue asiatiche, collegabile a una rete in continua espansione di informazioni, sia a livello nazionale sia internazionale, in cooperazione con progetti già attivi o in corso di realizzazione. L'obiettivo finale è creare uno spazio virtuale utilizzabile non solo dagli specialisti per fini accademici, ma anche da un pubblico più ampio per scopi educativi e divulgativi. Sebbene il progetto sia condotto in modo scientifico da esperti provenienti da diversi settori accademici, si può inoltre prevedere un significativo impatto didattico, che coinvolga numerosi studenti nello studio e nell'uso di questi strumenti digitali. L'insegnamento interdisciplinare, infatti, accresce l'apprendimento degli studenti e incoraggia un approccio integrato tra scuola e università. In particolare, siamo orgogliosi di sottolineare che il progetto

⁵ I dati disponibili sono per esempio stati utilizzati per le ricerche sui nomi di persona reali e fittizi o mitologici presenti nel *Confucius Sinarum Philosophus*, sui quali è in corso di pubblicazione in *Digital Classics Online* il seguente contributo: Balbo Andrea e Elisa Della Calce. *Detecting Eastern and Western Names in the Latin Corpus of the SERICA Project – With Special Regard to the Confucius Sinarum Philosophus (1687) as a Case Study*.

⁶ Per un progetto analogo si veda il CHCD della Boston University.

ha consentito l'impiego temporaneo di almeno venti giovani ricercatori, alcuni dei quali ancora studenti universitari o agli inizi della loro carriera accademica.

3. La continuazione di SERICA

Il progetto SERICA si è formalmente chiuso nel 2023, ma continua in maniera approfondita sia attraverso il progetto *Classica SERICA*, PRIN 2022, guidato da Andrea Balbo come P. I. in collaborazione con altre due unità, Siena, guidata da Francesco Stella e comprendente anche l'università di Macerata, Venezia, guidata da Tiziana Lippiello. Il progetto si conclude nel febbraio 2026 e ha dedicato un'attenzione particolare a opere quali il *Confucius Sinarum Philosophus* (1687) e ad autori come Cicerone, Quintiliano, Seneca, Livio e altri storici, con l'ulteriore obiettivo di mettere in luce il significato e il valore dei *Wertbegriffe* nelle opere neolatine, realizzando anche la preparazione e la marcatura dei testi latini destinati al sito web di SERICA. Una terza fase del progetto, che utilizzerà anche le risorse online di SERICA, consisterà nel progetto *SERICA Pedemontana*, consistente nella catalogazione di libri riguardanti l'Estremo Oriente, scritti in cinese o in latino tra il XVI e il XIX secolo, conservati in biblioteche e archivi pubblici e privati del Piemonte e della Valle d'Aosta. Ciò consentirà di mappare una porzione in larga parte trascurata delle antiche *provinciae Mediolanensis* e *Taurinensis* della Compagnia di Gesù, che nei territori oggetto dell'indagine mantenevano collegi, noviziati, residenze o parrocchie ad Alessandria, Aosta, Arona, Biella, Castelnuovo Scrivia, Chieri, Cuneo, Fenestrelle, Lanzo, Mondovì, Pinerolo, Saluzzo, Savigliano, Torino e Vercelli.

Questa indagine intende inoltre esaminare almeno due istituzioni diverse dai collegi e dagli enti gesuitici: l'Accademia delle Scienze di Torino, che sostiene il progetto e conserva una collezione di volumi e documenti in cinese e in latino (collegata al progetto *Chinese Collections*), e gli Archivi di Stato, presso i quali si ritiene plausibile siano custoditi documenti trasferiti da Genova e dalla Liguria, un tempo conservati presso il collegio gesuitico del capoluogo ligure. Nell'ambito del progetto rimangono attive le collane *Roma Sinica. Mutual interactions between Ancient Roman and Eastern Thought* di De Gruyter Brill, guidata da Andrea Balbo e Jaewon Ahn, Seoul National University, e *Documenta Orientalia*, codiretta da Andrea Balbo, Francesco Stella e Chiara Ombretta Tommasi. La prima ha già al suo attivo 6 volumi e il settimo, dal titolo *Along the Silk Roads*, è previsto nella primavera 2026; la seconda che conta 3 testi con altri in preparazione.

Nell'ambito di questi progetti si è creato ormai un forte *network* che unisce le università di Torino, Pisa, Venezia, Siena, Modena e Reggio Emilia, Enna, Fudan, Seoul e che sta sviluppando ulteriori connessioni con Roma La Sapienza e con Napoli L'Orientale. L'obiettivo sarà quello di avviare nuove progettualità su scala europea e non solo più italiana, sfruttando i circuiti ERC, Erasmus + e Horizon, ma rafforzando anche l'analisi territoriale della presenza di testi cinesi o che parlano della Cina nelle biblioteche italiane fino al XIX secolo. Essenziale sarà l'arricchimento del sito, che ha stretto accordi anche con progetti come *Textus Sinici*,⁷ e il miglioramento e il raffinamento degli strumenti di ricerca nell'ottica di una maggiore interoperabilità. L'ottica di SERICA è, quindi, quella dell'interazione continua, con l'obiettivo di costruire uno strumento di assoluto rilievo a livello internazionale.

⁷ <https://textus-sinici.org/>