Un LLM per le Scienze Umane e Sociali? ReSearch_SSH come esperimento europeo

ADAM FACI[1], ALESSIO MIASCHI[2], ANNE COMBE[3], PASCAL CUXAC[4], FRANCESCA FRONTINI[2], NICOLAS LARROUSSE[1], STÉPHANE POUYLLAU[1]

1 HUMA-NUM, CNRS, FRANCE - {NAME.SURNAME}@HUMA-NUM.FR
2 CNR-ILC, ITALY - {NAME.SURNAME}@ILC.CNR.IT
3 INRIA, FRANCE - ANNE.COMBE@INRIA.FR
4 INIST, CNRS, FRANCE - PASCAL.CUXAC@INIST.FR

ABSTRACT (ITALIANO)

L'adozione dei Large Language Models (LLM) nel contesto della ricerca scientifica, in particolare per attività di ricerca e scoperta bibliografica, solleva questioni rilevanti per le Scienze Umane e Sociali (SSH), soprattutto in relazione alla diversità delle pratiche disciplinari e alla valutazione dei risultati. Questo contributo presenta ReSearch_SSH, un caso d'uso sviluppato nel contesto del progetto europeo LLMs4EU e dell'infrastruttura ALT-EDIC, come esperimento di adattamento e valutazione di modelli linguistici orientato alle pratiche di ricerca nelle SSH e, più specificamente, nelle Digital Humanities. Il lavoro descrive l'architettura metodologica del caso d'uso, basata su strategie di *domain alignment*, *task-oriented fine-tuning* e su un approccio GraphRAG integrato con infrastrutture di scienza aperta come ISIDORE e ISTEX. Particolare attenzione è dedicata alle risorse dati multilingui, ai requisiti di fattualità e tracciabilità delle fonti e a un modello di valutazione qualitativa fondato sul coinvolgimento di un panel di esperti delle Digital Humanities. Il contributo si colloca nel dibattito delle Digital Humanities sugli LLM come componenti infrastrutturali per la ricerca, proponendo un approccio orientato alle pratiche disciplinari e alla partecipazione delle comunità scientifiche per guidare in modo critico lo sviluppo e la valutazione di sistemi di ricerca e scoperta basati su LLM in ambito SSH.
Parole chiave: Large Language Models; fine-tuning; valutazione qualitativa.

ABSTRACT (ENGLISH)

An LLM for the Social Sciences and Humanities? ReSearch_SSH as a European Experiment. The adoption of Large Language Models (LLMs) in the context of scientific research, particularly for bibliographic research and discovery, raises significant challenges for the Social Sciences and Humanities (SSH), especially with regard to the diversity of disciplinary practices and the evaluation of results. This paper presents ReSearch_SSH, a use case developed within the European project LLMs4EU and the ALT-EDIC infrastructure, as an experiment in adapting and evaluating language models grounded in SSH research practices, with a specific focus on Digital Humanities. The contribution describes the methodological architecture of the use case, based on *domain alignment*, *task-oriented fine-tuning*, and a GraphRAG approach integrated with open science infrastructures such as ISIDORE and ISTEX. Particular attention is devoted to multilingual data resources, requirements of factuality and source traceability, and a qualitative evaluation framework based on the involvement of a panel of Digital Humanities experts. The paper situates itself within the Digital Humanities debate on LLMs as infrastructural components for research, advocating a practice-oriented and community-driven approach to critically guide the development and evaluation of LLM-based research and discovery systems in the SSH domain.

Keywords: Large Language Models; fine-tuning; qualitative evaluation.

1. INTRODUZIONE

Negli ultimi anni, i Large Language Models (LLM) hanno trasformato profondamente il panorama delle tecnologie linguistiche e delle pratiche di ricerca nelle Digital Humanities (si vedano, tra i molti contributi possibili, Ciotti, 2023, e per un esempio di utilizzo degli LLM nelle Scienze Sociali e Umanistiche, Arachchige et al., 2025). Tuttavia, lo sviluppo di LLM e di piattaforme di *scientific discovery* a supporto della ricerca scientifica è oggi fortemente concentrato su lingue e contesti culturali dominanti e tende a privilegiare forme di produzione e validazione della conoscenza tipiche delle scienze sperimentali. Strumenti ampiamente diffusi per l'esplorazione della letteratura scientifica — come motori semantici,

sistemi di *question answering* o piattaforme di sintesi automatica basate su grandi corpora di articoli — si fondano prevalentemente su pubblicazioni in lingua inglese e su modelli epistemici centrati su articoli journal-based, metriche citazionali e risultati sperimentali. Questo orientamento, evidente in molte piattaforme di scoperta scientifica oggi disponibili, comporta il rischio di marginalizzare sia la diversità linguistica europea sia le specificità epistemiche e metodologiche della ricerca nelle Scienze Umane e Sociali, dove assumono rilievo altre tipologie di fonti, pratiche interpretative e forme di argomentazione. Come osservato da Fenlon (2017), infatti, la ricerca umanistica (ed in particolare le DH) ha visto di recente l'emergere di nuove forme di produzione scientifica, quali blog di ricerca, prodotti multimediali e grandi corpora digitali, che mettono in discussione le convenzioni tradizionali della pubblicazione accademica. In questo scenario si colloca il caso d'uso ReSearch_SSH che nasce dalla collaborazione tra infrastrutture SSH francesi e italiane nel contesto del progetto LLMs4EU (Large Language Models for the European Union), un progetto finanziato dalla Commissione Europea e coordinato da ALT-EDIC (Alliance for Language Technologies – European Digital Infrastructure Consortium[1]).
Alla luce di questo quadro, il contributo si articola come segue. Dopo aver inquadrato il tema dei Large Language Models a partire dalle pratiche di ricerca proprie delle Scienze Umane e Sociali, il lavoro presenta il progetto europeo LLMs4EU nel contesto della nuova infrastruttura ALT-EDIC, soffermandosi in particolare sul contributo e sul ruolo degli attori italiani. Viene quindi illustrato ReSearch_SSH come caso di studio di adattamento e fine-tuning di modelli linguistici orientato alle SSH, con attenzione alle risorse dati, all'architettura metodologica e alle strategie di valutazione adottate. Infine, il contributo discute il coinvolgimento delle comunità di ricerca — in particolare delle Digital Humanities — attraverso la costituzione di un panel di esperti, delineando possibili modalità di partecipazione della comunità AIUCD allo sviluppo e alla valutazione di LLM a supporto della ricerca.

2. PROGETTARE LLM A PARTIRE DALLE PRATICHE DELLE HUMANITIES

La ricerca nelle Scienze Umane e Sociali (SSH) è caratterizzata da pratiche interpretative, contestuali e riflessive, in cui dati, testi e metadati non costituiscono entità neutre, ma sono costruiti e reinterpretati nel tempo in relazione a quadri teorici, domande di ricerca e contesti storici e culturali. La letteratura sulle infrastrutture di ricerca per le *humanities* ha mostrato come tali pratiche richiedano ambienti digitali progettati a partire dalle esigenze delle comunità scientifiche: infrastrutture disciplinari come CLARIN ERIC, DARIAH ERIC, OPERAS o lo SSHOC Marketplace sono state concepite per sostenere l'uso condiviso di risorse digitali mantenendo un forte legame con la diversità disciplinare e linguistica (Branco et al., 2023; Dumouchel et al., 2020; König et al., 2023). Studi più recenti sulle infrastrutture di ricerca e sui servizi di archiviazione dati sottolineano inoltre come i dati umanistici e scientifici siano il risultato di pratiche comunitarie dinamiche, in cui archivi e service provider svolgono un ruolo attivo non solo di supporto, ma anche di mediazione e co-costruzione delle comunità di ricerca (Morselli et al., 2025).
Un ulteriore elemento distintivo della ricerca nelle humanities è la pluralità degli output scientifici e la centralità del multilinguismo come dimensione epistemica e metodologica della ricerca, ampiamente discussa negli studi di Digital Humanities multilingui (Balula & Leão, 2021; Viola & Spence, 2024). Accanto agli articoli, la produzione di conoscenza comprende dataset, corpora, workflows, edizioni digitali e forme di scrittura scientifica come i blog di ricerca, la cui rilevanza epistemica e la necessità di pratiche adeguate di valorizzazione e citazione sono ampiamente documentate (Barbot et al., 2024; Mayeur, 2017). In questo quadro, l'integrazione di funzionalità supportate da LLM per la ricerca bibliografica e documentale nelle SSH non può limitarsi a un uso generalista: i modelli linguistici devono essere progettati e valutati in modo da rispecchiare le pratiche di ricerca proprie delle *humanities*, favorendo approcci interdisciplinari e transdisciplinari e facilitando i contatti tra le comunità scientifiche. Questi principi trovano una prima concretizzazione nello use case ReSearch_SSH, elaborato nel contesto del progetto LLMs4EU guidato dall'infrastruttura ALT-EDIC. Nelle sezioni seguenti presenteremo questo progetto che mira a coinvolgere le comunità SSH ed in particolare DH nel fine-tuning e valutazione di modelli linguistici.

[1] https://www.alt-edic.eu/ (cons. 7/2/2026)

3. ALT-EDIC E LLMS4EU

ALT-EDIC è un European Digital Infrastructure Consortium[2], ovvero una forma giuridica europea pensata per consentire a più Stati membri di sviluppare e gestire congiuntamente infrastrutture digitali strategiche di interesse comune. In questo quadro, ALT-EDIC mira a sostenere l'eccellenza scientifica europea nel settore delle tecnologie del linguaggio e a promuovere la diversità linguistica europea, seguendo un modello di cooperazione che coinvolge istituzioni pubbliche, industria, società civile e ricerca (da cui la stretta collaborazione con CLARIN ERIC, l'infrastruttura di ricerca per le risorse linguistiche). L'Italia partecipa ad ALT-EDIC; alla guida del consorzio nazionale ci sono il CNR-ILC (istituto che ospita anche CLARIN-IT), la Fondazione Bruno Kessler (FBK) e il CINECA, che fornisce anche l'infrastruttura di calcolo.
Costituitosi nel 2024, ALT-EDIC sta consolidando le sue attività anche attraverso la partecipazione a diversi progetti europei tra loro complementari: tra questi, una menzione importante va a OpenEuroLLM, orientato allo sviluppo di modelli linguistici aperti e multilingui, che rappresentano un tassello chiave della strategia europea per un'Intelligenza Artificiale aperta, trasparente e allineata ai valori dell'Unione.
Il presente contributo origina invece dalle attività di un secondo progetto, LLMs4EU, che mira a potenziare gli aspetti più applicativi degli LLM nei vari settori strategici, con un approccio fortemente orientato ai casi d'uso. Il progetto si articola in cinque domini applicativi (turismo, servizi pubblici, telecomunicazioni, energia e scienza) e si sviluppa attorno al fine-tuning e all'adattamento di modelli aperti a bisogni specifici.

4. LO USE CASE RESEARCH_SSH

ReSearch_SSH è uno dei casi d'uso del dominio *scienza* ed è dedicato al supporto alle pratiche di ricerca nelle Scienze Umane e Sociali (SSH). Mira allo sviluppo e alla sperimentazione di modelli linguistici generativi adattati a compiti avanzati di scoperta, interrogazione e sintesi della letteratura scientifica, con un'attenzione specifica ai requisiti di fattualità, tracciabilità delle fonti ed esplicabilità dei risultati.
Il caso d'uso si fonda su una stretta integrazione tra partner infrastrutturali chiave e risorse europee, in particolare CNRS (attraverso le infrastrutture Huma-Num e INIST/ISTEX) e CNR-ILC (CLARIN-IT). L'obiettivo è sostenere flussi di lavoro realistici della ricerca SSH – quali la costruzione, l'estensione e l'aggiornamento dello stato dell'arte, nonché l'interrogazione ragionata di insiemi documentali – attraverso modelli multilingui (il caso d'uso si concentra per ora su francese e italiano, oltre che inglese) integrati in una piattaforma di ricerca ampiamente utilizzata come ISIDORE.[3]
Dati e focalizzazione disciplinare. Il fine-tuning partirà da modelli aperti, quali Salamandra (Gonzalez-Agirre et al., 2025) e il già citato OpenEuroLLM, e avverrà inizialmente sul corpus ISTEX (de Salabert & Barreaux, 2020) nella sua componente SSH (3 milioni di documenti, in diverse lingue, tra cui inglese, francese e italiano, per un totale di 2 miliardi di token circa), che rappresenta una base ampia e multidisciplinare della produzione scientifica in questo ambito. Questo corpus, costituito da testi completi in formato XML/TEI e da metadati ricchi e standardizzati, sarà utilizzato principalmente per le attività di *domain alignment*, al fine di adattare i modelli al discorso scientifico e alla terminologia propria delle SSH nel loro complesso.
Su questa base generalista, si introduce una focalizzazione progressiva sulle Digital Humanities, attraverso l'integrazione di risorse mirate provenienti sia dal contesto francese sia da quello italiano. In particolare, verranno utilizzati contenuti scientifici e blog accademici della piattaforma Hypotheses e dati e metadati di ricerca provenienti dal repository dati Nakala.[4] In seguito, potranno essere inseriti anche altri dati, tra cui atti di conferenze e pubblicazioni rappresentative delle comunità DH italiane, tra cui i *proceedings* di AIUCD, la rivista Umanistica Digitale, gli atti della conferenza CLiC e l'*Italian Journal of Computational Linguistics*. Queste risorse, rilasciate con licenze aperte, consentiranno di rafforzare l'allineamento multilingue e interculturale tra pratiche di ricerca SSH francesi, italiane e anglofone, e di ancorare il modello a comunità disciplinari ben definite.
Approccio metodologico e fine-tuning. L'approccio proposto privilegia un'architettura di Knowledge Graph–based Retrieval-Augmented Generation (GraphRAG, si veda tra gli altri Edge et al., 2025), concepita per garantire che le risposte generate siano sempre esplicitamente fondate su documenti di

[2] https://digital-strategy.ec.europa.eu/en/policies/edic (cons. 7/2/2026)
[3] https://isidore.science/ (cons. 7/2/2026)
[4] https://hypotheses.org/ (cons. 7/2/2026) ; https://www.nakala.fr/ (cons. 7/2/2026)

riferimento. In generale, i metodi di Retrieval-Augmented Generation (RAG) combinano modelli generativi con un modulo di retrieval, che seleziona dinamicamente porzioni rilevanti di conoscenza da una base documentale esterna, mitigando fenomeni di allucinazione e migliorando l'aderenza fattuale delle risposte. GraphRAG estende questo paradigma introducendo una rappresentazione strutturata della conoscenza sotto forma di grafo, che consente di modellare esplicitamente relazioni tra entità e documenti e di supportare processi di recupero e inferenza più contestualizzati e multi-hop.
Il processo di adattamento del modello si articola in più fasi. La prima fase consiste nell'allineamento di dominio (*domain alignment*), realizzato tramite pre-training su corpora SSH in francese e italiano, con l'obiettivo di migliorare la competenza linguistica e terminologica del modello senza introdurre comportamenti orientati a compiti specifici. La seconda fase è dedicata al task-*oriented fine-tuning*, in cui il modello viene adattato a compiti di ricerca attraverso *instruction tuning* integrato direttamente in una pipeline che implementa anche GraphRAG. I compiti target includono: comprensione di query di ricerca espresse in linguaggio naturale; recupero di documenti scientifici rilevanti; organizzazione dei risultati secondo criteri propri delle rassegne di letteratura; generazione di sintesi brevi e contestualizzate; risposta a domande su insiemi documentali strutturati; confronto e collegamento tra pubblicazioni sulla base di temi, autori e riferimenti. L'uso di grafi di conoscenza – si potranno usare quelli di Wikidata e OpenAIRE, oltre ai metadati di ISIDORE e Nakala – consente di arricchire il *retrieval*, migliorare il linking tra documenti e rafforzare l'esplicabilità dei risultati. Meccanismi interni di ragionamento (ad esempio rappresentazioni intermedie strutturate) possono essere impiegati a supporto dell'interpretazione delle query e della coerenza delle risposte, senza essere esposti direttamente all'utente finale.
Deployment e valutazione. Il modello così specializzato sarà messo in opera su una versione della piattaforma ISIDORE, concepita come una sorta di "ISIDORE AI", in cui sperimentare in modo controllato funzionalità avanzate di assistenza alla ricerca. Il sistema sarà integrato con il database e i grafi di conoscenza di ISIDORE e consentirà agli utenti di formulare interrogazioni complesse e ottenere risultati organizzati, sintetizzati e corredati da riferimenti espliciti alle fonti. L'interrogazione multilingue potrà essere supportata anche dall'uso di un modello traduttore.
La valutazione del sistema non si limiterà a metriche automatiche, ma includerà una valutazione "a panel" con il coinvolgimento di comunità di ricerca DH francesi e italiane. I panel di esperti saranno chiamati a valutare la qualità, l'utilità e l'affidabilità dei risultati prodotti, contribuendo anche a processi di ottimizzazione delle preferenze, orientati a migliorare l'allineamento del modello alle pratiche e ai criteri di rilevanza propri delle Digital Humanities.

5. QUESTIONI LEGALI E ADERENZA ALLE NORMATIVE EUROPEE

ReSearch_SSH è progettato in piena aderenza al quadro normativo europeo in materia di Intelligenza Artificiale, protezione dei dati e diritto d'autore. In primo luogo, le attività di adattamento e messa in opera dei modelli tengono conto dei vincoli introdotti dall'AI Act (Regulation (EU) 2024/1689, 2024), con particolare attenzione ai requisiti di trasparenza, tracciabilità delle fonti, gestione dei rischi e supporto alla valutazione umana, aspetti particolarmente rilevanti nel contesto di sistemi di supporto alla ricerca scientifica. L'adozione di un'architettura GraphRAG e l'utilizzo in una sandbox sperimentale contribuiscono a mitigare i rischi associati all'uso di modelli generativi in contesti sensibili come quello accademico.
Dal punto di vista della protezione dei dati personali, il progetto opera nel rispetto del GDPR, in particolare per quanto riguarda i dati raccolti nell'ambito delle attività di valutazione basate su panel di esperti. Tali dati, che possono includere opinioni, preferenze e valutazioni qualitative riconducibili a persone identificabili, sono trattati come dati sensibili, secondo principi di minimizzazione, limitazione delle finalità e conservazione controllata, e utilizzati esclusivamente per finalità di validazione scientifica e miglioramento del sistema.
Un'attenzione specifica è inoltre dedicata alle questioni di copyright e licensing. Tutti i dataset impiegati sono analizzati rispetto alle condizioni di utilizzo previste dalle rispettive licenze, incluse licenze non commerciali e, in alcuni casi, accordi specifici di riservatezza (*Non-Disclosure Agreements*). In linea con il dibattito recente, il progetto assume che la disponibilità di contenuti con licenza aperta (ad esempio CC-BY) non implichi automaticamente la loro riutilizzabilità per il fine-tuning di modelli linguistici. Come discusso da Spichtinger (2026), esistono infatti tensioni strutturali tra l'ethos della scienza aperta — fondato su condivisione, riuso e apertura dei dati — e l'"appetito" dei sistemi di Intelligenza Artificiale per grandi quantità di dati, che solleva interrogativi giuridici ed etici sul riuso dei contenuti scientifici al di là

delle finalità originarie per cui sono stati resi disponibili[5]. Ne consegue la necessità di una valutazione caso per caso delle condizioni di riutilizzo, incluse le eccezioni per il text and data mining previste dalla normativa europea.
In questo contesto, LLMs4EU adotta un approccio strutturato alla conformità legale ed etica: per ciascun caso d'uso sono previsti un Data Plan e un Legal and Ethics Compliance Plan, e il progetto include un intero Work Package dedicato alle questioni legali ed etiche. Questo assetto garantisce che la selezione dei dati, le strategie di fine-tuning e le modalità di deployment siano costantemente valutate alla luce delle normative europee vigenti e dei valori di responsabilità, affidabilità e sostenibilità che guidano lo sviluppo delle tecnologie linguistiche in ambito europeo.

6. VERSO UN PANEL DI ESPERTI

Un elemento centrale di ReSearch_SSH è la costituzione di un panel di esperti, composto da studiosi e professionisti delle Scienze Umane e Sociali, con un focus specifico sulle Digital Humanities. Questa scelta non è casuale: le DH rappresentano infatti un ambito delle SSH in cui l'uso di infrastrutture digitali, la riflessione metodologica e la sperimentazione di strumenti computazionali sono già parte integrante delle pratiche di ricerca. Il panel è quindi concepito non solo come strumento di valutazione qualitativa nelle fasi iniziali del progetto, ma come spazio di co-progettazione, in cui le comunità disciplinari partecipano attivamente alla definizione di scenari d'uso, requisiti metodologici e criteri di rilevanza. In questo senso future fasi di progettazione potranno utilizzare il feedback del panel per un ulteriore adattamento dei modelli.
Le attività di valutazione non riguarderanno esclusivamente gli output dei modelli, ma l'approccio del caso d'uso nel suo complesso, includendo aspetti quali la pertinenza, l'affidabilità e l'utilità dei risultati nei contesti di ricerca SSH, nonché l'individuazione di potenziali rischi epistemici, come allucinazioni, bias disciplinari o effetti di semplificazione. Il panel supporterà inoltre la sperimentazione di approcci di valutazione human-in-the-loop e di task che vadano oltre *benchmark* generalisti, privilegiando scenari complessi e situati, tipici della ricerca umanistica. Grazie all'integrazione con infrastrutture disciplinari per la scienza aperta e alla natura pubblica e federata degli strumenti sviluppati, ReSearch_SSH ambisce a creare le condizioni per un coinvolgimento duraturo delle comunità DH e SSH. In questa prospettiva, il panel non è concepito come un meccanismo limitato alla durata di LLMs4EU, ma come un modello di partecipazione e valutazione che possa essere mantenuto e fatto evolvere nel tempo all'interno delle infrastrutture di ricerca nazionali ed europee, anche oltre il ciclo di vita del progetto.

7. CONCLUSIONI

Nato nel contesto di LLMs4EU e dell'infrastruttura ALT-EDIC, lo use case ReSearch_SSH si configura come un laboratorio di sperimentazione per un uso responsabile e situato dei modelli linguistici generativi a supporto della ricerca nelle Scienze Umane e Sociali. Attraverso l'integrazione di strategie di adattamento dei modelli, architetture GraphRAG, infrastrutture di scienza aperta e processi di valutazione qualitativa basati su comunità di esperti, il caso d'uso propone un approccio che mette al centro l'impatto epistemico degli LLM sulle pratiche di scoperta, interpretazione e costruzione della conoscenza.
In una prospettiva di medio e lungo periodo, ReSearch_SSH apre a possibili collaborazioni strutturate con attori istituzionali e comunitari attivi nel panorama delle DH e delle SSH. In ambito italiano, un dialogo con il Dipartimento di Scienze Umane e Sociali, Patrimonio Culturale del CNR (CNR-DSU), e in particolare con il DSU Digital Humanities Center, potrebbe rafforzare l'integrazione tra sviluppo infrastrutturale, ricerca metodologica e comunità disciplinari. Analogamente, sul versante francese, il coinvolgimento del CNRS Humanities & Social Sciences (CNRS-INSHS) appare naturale per consolidare il legame tra strumenti di supporto alla ricerca, infrastrutture di scienza aperta e pratiche disciplinari condivise.
Un ruolo centrale è infine attribuito alle comunità scientifiche, in particolare ad AIUCD e alla corrispondente associazione francofona Humanistica, come spazi privilegiati per la definizione di scenari d'uso, criteri di valutazione e per una riflessione critica e continuativa sull'impatto dei modelli linguistici generativi nelle humanities. In questo senso, ReSearch_SSH ambisce a costituire non solo un risultato

[5] Dal punto di vista delle *Digital Humanities*, tale tensione non è un effetto collaterale, ma un nodo epistemologico e politico: la promessa di apertura, riuso e circolazione della conoscenza scientifica si confronta con pratiche di addestramento che rischiano di trasformare dati aperti in risorse opache, riaggregate e sottratte al controllo delle comunità che le hanno prodotte.

progettuale, ma un punto di partenza per un dialogo stabile tra infrastrutture, tecnologie e comunità di ricerca.

RINGRAZIAMENTI

Questo lavoro è stato sostenuto dal progetto LLMs4EU 'Large Language Models for the European Union', finanziato dall'Unione Europea tramite il Programma Europa Digitale (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) nell'ambito dell'accordo di sovvenzione 101198470.
Per la redazione dei Capitoli 4 e 5 di questo contributo è stata utilizzata IA generativa come supporto per la traduzione e la rielaborazione di passaggi originariamente redatti in lingua inglese.

BIBLIOGRAFIA

Arachchige, I. N., Frontini, F., Mitkov, R., & Rayson, P. (Eds). (2025). *Proceedings of the First on Natural Language Processing and Language Models for Digital Humanities*. INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2025.lm4dh-1

Balula, A., & Leão, D. (2021). Multilingualism within Scholarly Communication in SSH. A literature review. *JLIS.It*, *12*(2), 88–98. https://doi.org/10.4403/jlis.it-12672

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing Research Tools & Services Through Workflows in the SSH Open Marketplace. *Journal of Open Humanities Data*, *10*(1). https://doi.org/10.5334/johd.192

Branco, A., Eskevich, M., Frontini, F., Hajič, J., Hinrichs, E., Jong, F. de, Kamocki, P., König, A., Lindén, K., Navarretta, C., Piasecki, M., Piperidis, S., Pitkänen, O., Simov, K., Skadiņa, I., Trippel, T., Witt, A., & Zinn, C. (2023). The CLARIN infrastructure as an interoperable language technology platform for SSH and beyond. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-023-09658-z

Ciotti, F. (2023). Minerva e il pappagallo: IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *Testo e Senso*, (26), 289–315. https://doi.org/10.58015/2036-2293/671

de Salabert, C., & Barreaux, S. (2020). Vers un corpus optimal pour la fouille de textes: Stratégie de constitution de corpus spécialisés à partir d'ISTEX. In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, & S. Schneider (Eds), *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 4: Démonstrations et résumés d'articles internationaux* (pp. 66–69). ATALA. https://hal.science/hal-02768520

Dumouchel, S., Blotière, E., Barbot, L., Breitfuss, G., Chen, Y., Donato, F. D., Forbes, P., Petitfils, C., & Pohle, S. (2020). TRIPLE project: Building a discovery platform to enhance collaboration. *ITM Web of Conferences*, *33*, 03005. https://doi.org/10.1051/itmconf/20203303005

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2025). *From Local to Global: A Graph RAG Approach to Query-Focused Summarization* (arXiv:2404.16130). arXiv. https://doi.org/10.48550/arXiv.2404.16130

Fenlon, K. S. (2017). *Thematic research collections: Libraries and the evolution of alternative scholarly publishing in the humanities* [Text, University of Illinois at Urbana-Champaign]. https://hdl.handle.net/2142/99380

Gonzalez-Agirre, Aitor, et al. "Salamandra technical report." arXiv preprint arXiv:2502.08489 (2025).

König, A., Barbot, L., Grisot, C., Kurzmeier, M., & Gray, E. J. (2023). The SSH Open Marketplace and CLARIN. *Proceedings of the CLARIN Annual Conference*. https://doi.org/10.3384/ecp210006

Mayeur, I. (2017). Imparting Knowledge in Humanities. About Some Practices of Scientific Blogging on Hypothèses. In *Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices* (pp. 75–84). IOS Press. https://doi.org/10.3233/978-1-61499-769-6-75

Morselli, F., Touber, J., & Scharnhorst, A. (2025). *Fostering Data Communities—Perspective from a Data Archive Service Provider* (arXiv:2502.02321). arXiv. https://doi.org/10.48550/arXiv.2502.02321

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, L, 2024/1689.

Spichtinger, D. (2026). *Perspective Chapter: Fit for Purpose? Creative Commons Licensing for Research Data in the Age of Artificial Intelligence*. IntechOpen. https://doi.org/10.5772/intechopen.1013402

Viola, L., & Spence, P. (Eds). (2024). *Multilingual Digital Humanities*. Routledge.